

Compressive Embedding and Visualization using Graphs

Johan Paratte, Nathanaël Perraudin, Pierre Vandergheynst *

February 21, 2017

Abstract

Visualizing high-dimensional data has been a focus in data analysis communities for decades, which has led to the design of many algorithms, some of which are now considered references (such as t-SNE for example). In our era of overwhelming data volumes, the scalability of such methods have become more and more important. In this work, we present a method which allows to apply any visualization or embedding algorithm on very large datasets by considering only a fraction of the data as input and then extending the information to all data points using a graph encoding its global similarity. We show that in most cases, using only $\mathcal{O}(\log(N))$ samples is sufficient to diffuse the information to all N data points. In addition, we propose quantitative methods to measure the quality of embeddings and demonstrate the validity of our technique on both synthetic and real-world datasets.

Index terms— Graph signal processing, sampling, transductive learning, embedding, visualization

1 Introduction

DATA visualization is usually equivalent to mapping high-dimensional features in low dimension using distance preserving dimensionality reduction. This process, finding a low-dimensional embedding of high-dimensional data, has drawn a lot of attention from researchers in different fields.

Some methods are very fundamental such as Principle Component Analysis (PCA) or Linear Discriminant Analysis (LDA). Other well known methods use the hypothesis that the data can be well approximated by a low-dimensional

*EPFL, Ecole Polytechnique Fédérale de Lausanne, LTS2 Laboratoire de traitement du signal, CH-1015 Lausanne, Switzerland

manifold, such as Laplacian Eigenmaps [1], Isomap [2] or Local Linear Embedding (LLE) [3]. Another approach is to use a probabilistic model of both the high-dimensional and low-dimensional data distribution and optimize the distance preservation using the joint model. Examples of this approach are Stochastic Neighbor Embedding (SNE) [4] and its popular extension t-SNE [5] or LargeVis [6]. We refer the interested reader to this work [7], offering a comparative of numerous dimensionality reduction techniques.

From all those methods, two main pitfalls are the most prevalent. The first one is the lack of robustness to noisy real-world data and the second is bad scalability leading to unmanageable computing time for large datasets.

The first problem often arises when applying a global scheme which will work well on toy examples and fail on complex data, as the expected global model is only partially valid. A simple example would be the different results of Laplacian Eigenmaps which will yield the recovery of a perfect embedding for the Swissroll point cloud and poor results on large-scale complex and noisy data. This problem is traditionally mitigated by considering hypotheses on data to hold only locally, leading to techniques such as LLE, SNE and others.

The second, more important, issue of scalability is essential in today's world of ubiquitous and overwhelming data. It is even more crucial now that the increase in data creation cannot be well compensated by the physical limits unsettling Moore's law. Essentially, this fundamental issue of scalability is related to the notion of similarity. Indeed, the essential question one must be able to answer to represent data in low dimension is one of similarity : which data points are close to each other. This issue can be said to be fundamental because it naturally implies that the minimal complexity can only be super-linear, since one pass over each datapoint cannot be sufficient to infer a similarity matrix with a quadratic number of entries. Some of the popular methods mentioned above do have an intrinsic quadratic regime and parallelized or approximated variants that scale better, but at a cost. An illustrative example is t-SNE which is $\mathcal{O}(N^2)$ in its original implementation and is mostly used with an approximated and accelerated version (Barnes-Hut t-SNE [8]) in $\mathcal{O}(N \log(N))$.

As we saw, the two issues mentioned above are related to the concepts of locality and similarity. Expressing both notions naturally leads to the concept of a similarity graph whose edges link the closest points, weighted by the distance between them. This general idea is actually one of the most used tool when computing embeddings, either explicitly in methods such as Laplacian Eigenmaps or LargeVis, or implicitly, using probability distributions as random walk matrices (e.g. SNE). Of course, constructing a similarity graph has the same complexity issue as the one mentioned above. This is why approximated sparse nearest-neighbors graphs are often used in practice, as they can be computed very efficiently using Approximated Nearest Neighbors (ANN) techniques (e.g. FLANN [9]).

In this work, we propose a general framework for accelerating any embedding algorithm using a graph encoding the data similarity. Our technique is supported by modern tools of Graph Signal Processing allowing to use the graph at

both local and global scales. The main idea is to use only a subset of the data on which to apply an embedding algorithm and then diffuse the information using the graph. Our main contribution which we call Compressive Embedding (CE) is made possible by two complementary mechanisms : a graph sampling scheme to create the sketch and diffusion routines to extend the information on the sketch to all data points.

Contributions Below we summarize the main contributions of this work :

- graph sampling schemes and theorems stating the minimum number of samples necessary to capture energy everywhere
- transductive learning algorithms to extend the embedding information computed on the samples to all datapoints using localized low pass graph filters
- new quantitative measures of the quality of the visualizations based on graph cuts and localized filters
- experiments on synthetic and real data sets showing the superior scalability of this method compared to the state-of-the-art

Organization The paper is organized as follows. In Section 2, we recall the fundamentals of graph signal processing and define the notations. Section 3 develops the results on our sampling method based on the energy of localized kernels. Section 4 uses localized filters to define generalized metrics used in the following sections. Section 5 describes the different methods to extend the information from the sampled nodes to all data points. Section 7 describes our proposed methods to compute a quantitative measure of the quality of embeddings. In Section 8, we show the validity and benefits of our method and compare with the state-of-the-art through several experiments. Finally, Section 9 proposes interesting open problems in the domain as well as potential future work to address.

2 Background

Graph nomenclature Let us define $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ as an undirected weighted graph where \mathcal{V} is the set of vertices and \mathcal{E} the set of edges representing connections between nodes in \mathcal{V} . The vertices $v \in \mathcal{V}$ of the graph are ordered from 1 to $N = |\mathcal{V}|$. The matrix \mathbf{W} , which is symmetric and positive, is called the weighted adjacency matrix of the graph \mathcal{G} . The weight \mathbf{W}_{ij} represents the weight of the edge between vertices v_i and v_j and a value of 0 means that the two vertices are not connected. The degree $d(i)$ of a node v_i is defined as the sum of the weights of all its edges $d(i) = \sum_{j=1}^N \mathbf{W}_{ij}$. Finally, a graph signal is defined as a vector of scalar values over the set of vertices \mathcal{V} where the i -th component of the vector is the value of the signal at vertex v_i .

Spectral theory The combinatorial Laplacian operator \mathbf{L} can be defined from the weighted adjacency matrix as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ with \mathbf{D} being the degree matrix defined as a diagonal matrix with $D_{ii} = d(i)$. One alternative and often used Laplacian definition is the normalized Laplacian $\mathbf{L}_n = \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{\frac{1}{2}}$. Since the weight matrix \mathbf{W} is symmetric positive semi-definite, so is \mathbf{L} by construction. By application of the spectral theorem, we know that \mathbf{L} can be decomposed into an orthonormal basis of eigenvectors noted $\{\mathbf{u}_\ell\}_{\ell=0,1,\dots,N-1}$. The ordering of the eigenvectors is given by the eigenvalues noted $\{\lambda_\ell\}_{\ell=0,1,\dots,N-1}$ sorted in ascending order $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{N-1} = \lambda_{\max}$. In a matrix form we can write this decomposition as $\mathbf{L} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^*$ with $\mathbf{U} = (\mathbf{u}_1 | \mathbf{u}_2 | \dots | \mathbf{u}_{N-1})$ the matrix of eigenvectors and $\mathbf{\Lambda}$ the diagonal matrix containing the eigenvalues in ascending order. Given a graph signal \mathbf{x} , its graph Fourier transform is thus defined as $\hat{\mathbf{x}} = \mathcal{F}(\mathbf{x}) = \mathbf{U}^* \mathbf{x}$, and the inverse transform $\mathbf{x} = \mathcal{F}^{-1}(\hat{\mathbf{x}}) = \mathbf{U} \hat{\mathbf{x}}$. It is called a Fourier transform by analogy to the continuous Laplacian whose spectral components are Fourier modes, and the matrix \mathbf{U} is sometimes referred to as the graph Fourier matrix (see e.g., [10]). By the same analogy, the set $\{\sqrt{\lambda_\ell}\}_{\ell=0,1,\dots,N-1}$ is often seen as the set of graph frequencies [11].

Graph filtering In traditional signal processing, filtering can be carried out by a pointwise multiplication in Fourier. Thus, since the graph Fourier transform is defined, it is natural to consider a filtering operation on the graph using a multiplication in the graph Fourier domain. To this end, we define a graph filter as a continuous function $g : \mathbb{R}_+ \rightarrow \mathbb{R}$ directly in the graph Fourier domain. If we consider the filtering of a signal \mathbf{x} , whose graph Fourier transform is written $\hat{\mathbf{x}}$, by a filter g the operation in the spectral domain is a simple multiplication $\hat{\mathbf{x}}'[\ell] = g(\lambda_\ell) \cdot \hat{\mathbf{x}}[\ell]$, with \mathbf{x}' and $\hat{\mathbf{x}}'$ the filtered signal and its graph Fourier transform respectively. Using the graph Fourier matrix to recover the vertex-based signals we get the explicit matrix formulation for graph filtering:

$$\mathbf{x}' = \mathbf{U} g(\mathbf{\Lambda}) \mathbf{U}^* \mathbf{x},$$

where $g(\mathbf{\Lambda}) = \text{diag}(g(\lambda_0), g(\lambda_1), \dots, g(\lambda_{N-1}))$. The graph filtering operator $g(\mathbf{L}) := \mathbf{U} g(\mathbf{\Lambda}) \mathbf{U}^*$ is often used to reformulate the graph filtering equation as a simple vector-matrix operation $\mathbf{x}' = g(\mathbf{L}) \mathbf{x}$.

Since the filtering equation defined above involves the full set of eigenvectors \mathbf{U} , it implies the diagonalization of the Laplacian \mathbf{L} which is costly for large graphs. To circumvent this problem, one can represent the filter g as a polynomial approximation, since polynomial filtering only involves the multiplication of the signal by a power of \mathbf{L} of the same order as the polynomial. Filtering using good polynomial approximations can be done using Chebyshev or Lanczos polynomials [12, 13].

Localization operator The concept of translation, which is well defined in traditional signal processing cannot be directly applied to graphs, as they can be irregular. However, inspired by the notion of translation, we can define the localization of a function g defined on the graph spectrum as a convolution

with a Kronecker delta $\widehat{\mathcal{T}_i g[\ell]} = g(\lambda_\ell) \cdot \hat{\delta}_i = g(\lambda_\ell) \cdot \mathbf{u}_\ell[i]$, where \mathcal{T} is called the localization operator, and \mathcal{T}_i means localization at vertex i . Going back to the vertex domain, we get :

$$\mathcal{T}_i g[n] = \mathcal{F}^{-1} \left(g \cdot \hat{\delta}_i \right) [n] = \sum_{\ell=0}^{N-1} g(\lambda_\ell) \mathbf{u}_\ell^*[i] \mathbf{u}_\ell[n] = (g(\mathbf{L}))_{in}.$$

The reason for calling \mathcal{T}_i a localization operator comes from the fact that for smooth functions g , $\mathcal{T}_i g$ is localized around the vertex i . The proof of this result and more information on the localization operator can be found in [14]. The localization of filters is quite naturally called atoms as a filtering operation of a signal \mathbf{x} using a filter g can be expressed as $\mathbf{x}'[i] = \langle \mathbf{x}, \mathcal{T}_i g \rangle$.

Additional notation We use $\|\mathbf{A}\|_{op} = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$ for the induced norm of the matrix \mathbf{A} and $\|\mathbf{A}\|_F = \sqrt{\sum_i \sum_j \mathbf{A}_{ij}^2}$ for the Froebenius norm. The maximum eigenvalue of a matrix is written $\sigma_{\max}(\mathbf{A})$.

We reserve the number notation for vectors. For example, we write the ℓ_2 Euclidean norm as $\|\mathbf{x}\|_2 = \sqrt{\sum_i \mathbf{x}_i^2}$ and the ℓ_∞ uniform (sup) norm $\|\mathbf{x}\|_\infty = \max_i |\mathbf{x}_i|$. We abusively use the ℓ_0 to count the number of non-zero elements in a vector. Furthermore, when an univariate function g is applied to a vector $\boldsymbol{\lambda}$, we mean $[g(\boldsymbol{\lambda})]_i = g(\lambda_i)$. As a result, $\|g(\boldsymbol{\lambda})\|_0 = k$ is the number of eigenvalues where $g(\lambda_\ell) \neq 0$.

Given a kernel g , we define \mathbf{U}_k as a $N \times k$ matrix made of the k columns of \mathbf{U} where $g(\lambda_\ell) \neq 0$. Similarly, we denote $\boldsymbol{\Lambda}_k$ the $k \times k$ diagonal matrix containing the associated eigenvalues. Note that we have

$$g(\mathbf{L}) = \mathbf{U} g(\boldsymbol{\Lambda}) \mathbf{U}^* = \mathbf{U}_k g(\boldsymbol{\Lambda}_k) \mathbf{U}_k^* = \mathbf{U}_k \mathbf{U}_k^* g(\mathbf{L}).$$

3 Random sampling on graphs

In this section, we first define a graph sampling schemes and then prove related theoretical limits. In particular, it is of particular interest to understand the number of samples needed in order to diffuse energy on every node by localizing filters on the samples. We will prove that the number of samples needed is directly linked with the rank of the filter.

3.1 Adaptive sampling scheme

Let us define the probability distribution \mathcal{P} represented by a vector $\mathbf{p} \in \mathbb{R}^N$. We use two different sampling schemes. Uniform sampling is given by the prob-

ability vector

$$\mathbf{p}_i = \frac{1}{N},$$

and adapted sampling is given by

$$\mathbf{p}_i = \frac{\|\mathcal{T}_i g\|_2^2}{\|g(\boldsymbol{\lambda})\|_2^2}.$$

Remember that we have $\sum_i \|\mathcal{T}_i g\|_2^2 = \|g\|_2^2$, implying that $\sum_i p_i = 1$. Let us associate the matrix

$$P := \text{diag}(p) \in \mathbb{R}^{N \times N}$$

to p .

Then, we draw independently (with replacement) M indices $\Omega := \{\omega_1, \dots, \omega_M\}$ from the set $\{1, \dots, N\}$ according to the probability distribution \mathbf{p} . We have

$$\mathbb{P}[\omega_j = i] = \mathbf{p}_i, \quad \forall i \in \{1, \dots, N\}, \quad \forall j \in \{1, \dots, M\}.$$

For any signal $\mathbf{x} \in \mathbb{R}^N$, defined on the vertices of the graph, its sampled version $\mathbf{y} \in \mathbb{R}^M$ satisfies

$$\mathbf{y}_j := \mathbf{x}_{\omega_j} \quad \forall j \in \{1, \dots, M\}.$$

Finally, the downsampling matrix $M \in \mathbb{R}^{M \times N}$ is defined as

$$M_{ij} = \begin{cases} 1 & \text{if } i = \omega_j \\ 0 & \text{otherwise,} \end{cases}$$

for all $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, M\}$. Note that $\mathbf{y} = M\mathbf{x}$.

3.2 Embedding Theorems

The first theorem shows that given enough samples, the random projection $MP^{-\frac{1}{2}}g(\mathbf{L})\mathbf{x}$ conserves the energy contained in $g(\mathbf{L})\mathbf{x}$. In this sense, given enough samples, it is an embedding of $g(\mathbf{L})\mathbf{x}$.

Theorem 1. *Given a graph \mathcal{G} and a kernel g with a given rank $\|g(\boldsymbol{\lambda})\|_0 = k$, given $\delta > 0$ and using the sampling scheme of Section 3.1, if*

$$M \geq 2 \frac{1}{\delta^2} \frac{\|g(\boldsymbol{\lambda})\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} \left(1 + \frac{\delta}{3}\right) \log \left(\frac{2k}{\epsilon}\right)$$

we have with a probability of $1 - \epsilon$ for all \mathbf{x} :

$$\left| \frac{\frac{1}{M} \left\| MP^{-\frac{1}{2}}g(\mathbf{L})\mathbf{x} \right\|_2^2 - \|g(\mathbf{L})\mathbf{x}\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} \right| \leq \delta \|U_k^* \mathbf{x}\|_2^2 \leq \delta \|\mathbf{x}\|_2^2. \quad (1)$$

Note that the above expression is normalized by $\|g(\boldsymbol{\lambda})\|_\infty^2$ in order to remove the scaling factor of the kernel g .

Let us now analyze the most important term of the bound:

$$\frac{\|g(\boldsymbol{\lambda})\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} = \frac{\sum_\ell g^2(\lambda_\ell)}{\max_\ell g^2(\lambda_\ell)}. \quad (2)$$

It is a measure of concentration of the kernel on its support. It is maximized with the value $\frac{\|g(\boldsymbol{\lambda})\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} = k$ when g is a rectangle. In general, it will be small for concentrated kernels. For example, a rapidly decreasing kernel such as the heat kernel ($g(x) = e^{-x^\tau}$) will lead to a very small ratio.

Note that contrarily to almost all bound available in the literature this bound does not require the kernel to be low rank but only concentrated. For a comparison [15, Corollary 2.3] requires

$$M \geq \frac{3}{\delta^2} k \log \left(\frac{2k}{\epsilon} \right).$$

Optimality of the sampling scheme. Although we have no formal proof of optimality, the sampling scheme presented in Section 3.1 is a good candidate. Indeed, when reading the proof of Theorem 1, the reader may notice that it minimizes the number of samples M .

Building on top of Theorem 1, we establish a lower bound on the number of samples required by Algorithm 1 to capture enough information from each node with a given confidence level. It will ensure that the information diffused from the samples can reach all nodes.

Theorem 2. *Using the sampling scheme described in Section 3.1, for $\delta > 0$, a graph \mathcal{G} and a kernel g such that $\|g(\boldsymbol{\lambda})\|_0 = k$, each node i is guaranteed with a probability $1 - \epsilon$ to have*

$$\frac{\frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{-\frac{1}{2}} \mathcal{T}_i g \right\|_2^2}{\|\mathcal{T}_i g\|_2^2} \geq 1 - \delta,$$

given that the number of samples satisfies

$$M \geq \frac{2a}{\delta^2} \left(1 + \frac{\delta}{3} \right) \log \left(\frac{k}{\epsilon} \right),$$

$$\text{where } a = \frac{\|g(\boldsymbol{\lambda})\|_2^2 \|g(\boldsymbol{\lambda})\|_\infty^2 \|\mathbf{U}_k^* \boldsymbol{\delta}_i\|_2^4}{\|\mathcal{T}_i g\|_2^4}.$$

Theorem 2 warrants that given enough samples M , Algorithm 1 captures with some probability $1 - \epsilon$ (close to 1), at least a good percentage of the energy at node i . The factor a is always greater than 1 and varies depending on the shape of the kernel g and of the graph eigenvectors. However it is $\mathcal{O}(k)$ and exactly

equal to k if g is a rectangular kernel. Indeed, a simple transformation shows that

$$a = \frac{\|g(\boldsymbol{\lambda})\|_2^2 \|g(\boldsymbol{\lambda})\|_\infty^2 \|\mathbf{U}_k^* \boldsymbol{\delta}_i\|_2^4}{\|\mathcal{T}_i g\|_2^4} = \frac{\sum_\ell g^2(\lambda_\ell)}{\max_\ell |g^2(\lambda_\ell)|} \left(\frac{\max_\ell |g^2(\lambda_\ell)| \sum_{\ell \in \mathcal{K}} \mathbf{u}_\ell^2[i]}{\sum_\ell g^2(\lambda_\ell) \mathbf{u}_\ell^2[i]} \right)^2.$$

The first term is smaller than k but is usually close to k for a kernel close to a rectangle. The second term is greater than 1 but close to 1 given that the kernel is close to a rectangle.

Problematically, this bound becomes loose if the kernel g has a large rank because of the term $\sum_{\ell \in \mathcal{K}} \mathbf{u}_\ell^2[i]$. To cope with this problem we can use another kernel g' that is a low-rank approximation of g .

Theorem 3. *Given a graph \mathcal{G} , let g' (with $\|g'(\boldsymbol{\lambda})\|_0 = k$) to be the rank k approximation of the kernel g , i.e.,*

$$g(\lambda_\ell) = \begin{cases} g'(\lambda_\ell) & \text{for the } k \text{ greatest values of } |g(\lambda_\ell)| \\ 0 & \text{otherwise.} \end{cases}$$

Using the sampling scheme described in Section 3.1 with the kernel g , for $\delta > 0$, each node i is assured with a probability $1 - \epsilon$ to have

$$\frac{\frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{\frac{1}{2}} \mathcal{T}_i g \right\|_2^2}{\|\mathcal{T}_i g\|_2^2} \geq 1 - \delta - \frac{\|\mathcal{T}_i(|g'| - |g|)\|_2^2}{\|\mathcal{T}_i g\|_2^2}$$

providing the number of samples satisfies¹

$$M \geq 2 \frac{1}{\delta^2} \frac{\|g'(\boldsymbol{\lambda})\|_2^2 \|g'(\boldsymbol{\lambda})\|_\infty^2 \|\mathbf{U}_k^* \boldsymbol{\delta}_i\|_2^4}{\|\mathcal{T}_i g\|_2^4} \left(1 + \frac{\delta}{3} \right) \log \left(\frac{k}{\epsilon} \right).$$

Using Theorem 3, the number of samples M required can be highly reduced. Indeed, when the kernel g is well concentrated but not low rank, we trade some approximation error encoded by $\frac{\|\mathcal{T}_i(|g'| - |g|)\|_2^2}{\|\mathcal{T}_i g\|_2^2}$ (which will be low if g is concentrated) but we will need a smaller number of samples due to the fact that g' is low rank. This theorem can be interesting for a heat kernel for example.

4 Metrics based on localized filters

Before moving on to the information diffusion from the samples, we need to take a closer look to localized filters and in particular see how they can be used to measure distances or correlations between nodes.

¹Note that $\|\mathcal{T}_i g\|_2^2 \geq \|\mathcal{T}_i g'\|_2^2$.

4.1 Localized Kernel Distance

Since localized filters are proven to be concentrated in the vertex domain (see [11, Theorem 1]), it seems natural to use them to get geodesic measures or correlations between nodes. To this end, we introduce the Localized Kernel Distance (LKD), which is defined as :

$$\text{LKD}(i, j) = 1 - \frac{\mathcal{T}_i g^2[j]}{\|\mathcal{T}_i g\| \|\mathcal{T}_j g\|}. \quad (3)$$

Let us now examine its properties by stating the following theorem:

Theorem 4. *The space $(\mathcal{V}, \text{LKD})$ with \mathcal{V} the vertex set of a graph and LKD as defined in 3 is a pseudosemimetric space, that is, for every $x, y \in \mathcal{V}$:*

1. $\text{LKD}(x, y) \geq 0$
2. $\text{LKD}(x, x) = 0$
3. $\text{LKD}(x, y) = \text{LKD}(y, x)$

Proof. First, let us derive an alternative form of (3) :

$$\text{LKD}(x, y) = 1 - \frac{\langle \mathcal{T}_x g, \mathcal{T}_y g \rangle}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \quad (4)$$

This can be derived as follows :

$$\begin{aligned} \text{LKD}(x, y) &= 1 - \frac{\mathcal{T}_x g^2[y]}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= 1 - \frac{\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^*[x] \mathbf{u}_{\ell}[y]}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= 1 - \frac{\sum_{\ell} (g(\lambda_{\ell}) \mathbf{u}_{\ell}^*[x]) (g(\lambda_{\ell}) \mathbf{u}_{\ell}[y])}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= 1 - \frac{\sum_{\ell} (g(\lambda_{\ell}) \mathbf{u}_{\ell}^*[x]) (g(\lambda_{\ell}) \mathbf{u}_{\ell}^*[y]) \sum_n \mathbf{u}_{\ell}[n]^2}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= 1 - \frac{\sum_n \sum_{\ell} (g(\lambda_{\ell}) \mathbf{u}_{\ell}^*[x] \mathbf{u}_{\ell}[n]) (g(\lambda_{\ell}) \mathbf{u}_{\ell}^*[y] \mathbf{u}_{\ell}[n])}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= 1 - \frac{\langle \mathcal{T}_x g, \mathcal{T}_y g \rangle}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \end{aligned}$$

Now let us verify the properties one by one :

1. We have using (4) :

$$\begin{aligned}\text{LKD}(x, y) &= 1 - \frac{\langle \mathcal{T}_x g, \mathcal{T}_y g \rangle}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &\geq 0\end{aligned}$$

where the last inequality stands because $\langle \mathcal{T}_x g, \mathcal{T}_y g \rangle \leq \|\mathcal{T}_x g\| \|\mathcal{T}_y g\|$ (Cauchy-Schwartz inequality).

2. Let us verify that $x = y \Rightarrow \text{LKD}(x, y) = 0$:

$$\begin{aligned}\text{LKD}(x, y) &= \text{LKD}(x, x) \\ &= 1 - \frac{\mathcal{T}_x g^2[x]}{\|\mathcal{T}_x g\| \|\mathcal{T}_x g\|} \\ &= 1 - \frac{\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^*[x] \mathbf{u}_{\ell}[x]}{\|\mathcal{T}_x g\|^2} \\ &= 1 - \frac{\sum_{\ell} (g(\lambda_{\ell}) \mathbf{u}_{\ell}[x])^2 \sum_n \mathbf{u}_{\ell}[n]^2}{\|\mathcal{T}_x g\|^2} \\ &= 1 - \frac{\sum_n \sum_{\ell} (g(\lambda_{\ell}) \mathbf{u}_{\ell}[x] \mathbf{u}_{\ell}[n])^2}{\|\mathcal{T}_x g\|^2} \\ &= 1 - \frac{\|\mathcal{T}_x g\|^2}{\|\mathcal{T}_x g\|^2} \\ &= 0\end{aligned}$$

3. Finally, we have

$$\begin{aligned}\text{LKD}(x, y) &= 1 - \frac{\mathcal{T}_x g^2[y]}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= 1 - \frac{\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^*[x] \mathbf{u}_{\ell}[y]}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= 1 - \frac{\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^*[y] \mathbf{u}_{\ell}[x]}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= 1 - \frac{\mathcal{T}_y g^2[x]}{\|\mathcal{T}_x g\| \|\mathcal{T}_y g\|} \\ &= \text{LKD}(y, x)\end{aligned}$$

□

Theorem 5. *The space $(\mathcal{V}, \text{LKD})$ with \mathcal{V} the vertex set of a graph and LKD as defined in 3, with g constant, is a semimetric space, that is, for every $x, y \in \mathcal{V}$:*

1. $\text{LKD}(x, y) \geq 0$
2. $\text{LKD}(x, y) = 0 \Leftrightarrow x = y$
3. $\text{LKD}(x, y) = \text{LKD}(y, x)$

Proof. Properties 1 and 3, as well as the backward implication are still valid as stated in Theorem 4.

Now let us check that $\text{LKD}(x, y) = 0 \Rightarrow x = y$.

We want to do it by contradiction and thus search any x, y , $x \neq y$ for which $\text{LKD}(x, y) = 0$, implying :

$$\langle \mathcal{T}_x g, \mathcal{T}_y g \rangle = \|\mathcal{T}_x g\| \|\mathcal{T}_y g\| \quad (5)$$

We can rewrite this equality as :

$$\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^*[x] \mathbf{u}_{\ell}[y] = \sqrt{\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^2[x]} \sqrt{\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^2[y]} \quad (6)$$

For $g(x) = c$, with $c > 0$ a constant, the left hand side is :

$$\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^*[x] \mathbf{u}_{\ell}[y] = c^2 \sum_{\ell} \mathbf{u}_{\ell}^*[x] \mathbf{u}_{\ell}[y] = 0 \quad (7)$$

The last equality comes from the fact that two lines of an orthonormal matrix are orthogonal, and $x \neq y$.

Now the right-hand side is :

$$\sqrt{\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^2[x]} \sqrt{\sum_{\ell} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}^2[y]} = c^2 \sum_{\ell} \mathbf{u}_{\ell}^2[x] \sum_{\ell} \mathbf{u}_{\ell}^2[y] = c^2 \quad (8)$$

with the last equality coming from the fact that \mathbf{U} is an orthonormal basis.

Now, since $0 \neq c^2$ we have a contradiction, and thus the proof is completed. \square

4.2 Kernelized Diffusion Distance

Another approach to use localized atoms to define distances is to measure the norm of the difference between a filter localized at two different nodes. We call it the Kernelized Diffusion Distance and define it as:

$$\text{KDD}(i, j) = \|\mathcal{T}_i g - \mathcal{T}_j g\|, \quad (9)$$

where g is a kernel defined in the graph spectral domain. Before going further, and as it will be useful later, let us derive a corollary definition of 9 :

$$\text{KDD}(i, j) = \sqrt{\sum_{\ell} g(\lambda_{\ell})^2 (\mathbf{u}_{\ell}^*[i] - \mathbf{u}_{\ell}^*[j])^2}. \quad (10)$$

This alternative definition can be quickly derived as follows :

$$\begin{aligned} \text{KDD}(i, j)^2 &= \|\mathcal{T}_i g - \mathcal{T}_j g\|^2 \\ &= \sum_n \left(\sum_{\ell} g(\lambda_{\ell}) \mathbf{u}_{\ell}^*[i] \mathbf{u}_{\ell}[n] - \sum_{\ell} g(\lambda_{\ell}) \mathbf{u}_{\ell}^*[j] \mathbf{u}_{\ell}[n] \right)^2 \\ &= \sum_n \left(\sum_{\ell} g(\lambda_{\ell}) (\mathbf{u}_{\ell}^*[i] - \mathbf{u}_{\ell}^*[j]) \mathbf{u}_{\ell}[n] \right)^2 \\ &= \sum_n \sum_{\ell} g(\lambda_{\ell})^2 (\mathbf{u}_{\ell}^*[i] - \mathbf{u}_{\ell}^*[j])^2 \mathbf{u}_{\ell}^2[n] \\ &= \sum_{\ell} g(\lambda_{\ell})^2 (\mathbf{u}_{\ell}^*[i] - \mathbf{u}_{\ell}^*[j])^2 \sum_n \mathbf{u}_{\ell}^2[n] \\ &= \sum_{\ell} g(\lambda_{\ell})^2 (\mathbf{u}_{\ell}^*[i] - \mathbf{u}_{\ell}^*[j])^2 \end{aligned}$$

which implies 10 by taking the square root on both sides.

Let us now examine the properties of the KDD by stating the following theorem:

Theorem 6. *The space $(\mathcal{V}, \text{KDD})$ with \mathcal{V} the vertex set of a graph and KDD as defined in 9 is a pseudometric space, that is, for every $x, y, z \in \mathcal{V}$:*

1. $\text{KDD}(x, y) \geq 0$
2. $\text{KDD}(x, y) = \text{KDD}(y, x)$
3. $\text{KDD}(x, z) \leq \text{KDD}(x, y) + \text{KDD}(y, z)$

Proof. Let us verify the properties in order :

1. This property holds trivially due to the positivity of the norm $\|\cdot\|$.
2. We have

$$\begin{aligned} \text{KDD}(x, y) &= \|\mathcal{T}_x g - \mathcal{T}_y g\| \\ &= \sqrt{\sum_{\ell} g(\lambda_{\ell})^2 (\mathbf{u}_{\ell}^*[x] - \mathbf{u}_{\ell}^*[y])^2} \\ &= \sqrt{\sum_{\ell} g(\lambda_{\ell})^2 (\mathbf{u}_{\ell}^*[y] - \mathbf{u}_{\ell}^*[x])^2} \\ &= \|\mathcal{T}_y g - \mathcal{T}_x g\| \\ &= \text{KDD}(y, x) \end{aligned}$$

3. We have

$$\begin{aligned}
\text{KDD}(x, z) &= \|\mathcal{T}_x g - \mathcal{T}_z g\| \\
&= \|\mathcal{T}_x g - \mathcal{T}_y g + \mathcal{T}_y g - \mathcal{T}_z g\| \\
&\leq \|\mathcal{T}_x g - \mathcal{T}_y g\| + \|\mathcal{T}_y g - \mathcal{T}_z g\| \\
&= \text{KDD}(x, y) + \text{KDD}(y, z)
\end{aligned}$$

which holds using the triangle inequality for vectors.

□

Now that we proved that the KDD is a pseudo-metric, we only need to have the identity of the indiscernibles, i.e. $\text{KDD}(i, j) = 0 \Leftrightarrow i = j$ to prove it is a metric. However, we can only do it using an additional hypothesis on g . This is formulated in the following theorem :

Theorem 7. *The space $(\mathcal{V}, \text{KDD})$ with \mathcal{V} the vertex set of a graph and KDD as defined in 9, with g being full rank, is a metric space, that is, for every $x, y, z \in \mathcal{V}$:*

1. $\text{KDD}(x, y) \geq 0$
2. $\text{KDD}(x, y) = \text{KDD}(y, x)$
3. $\text{KDD}(x, z) \leq \text{KDD}(x, y) + \text{KDD}(y, z)$
4. $\text{KDD}(x, y) = 0 \Leftrightarrow x = y$

Proof. Properties 1-3 are still valid as stated in Theorem 6.

Now let us check Property 4.

- We first prove $x = y \Rightarrow \text{KDD}(x, y) = 0$:

$$\begin{aligned}
d_g(x, y) &= d_g(x, x) \\
&= \|\mathcal{T}_x g - \mathcal{T}_x g\| \\
&= \sqrt{\sum_{\ell} g(\lambda_{\ell})^2 (\mathbf{u}_{\ell}^*[x] - \mathbf{u}_{\ell}^*[x])^2} \\
&= 0
\end{aligned}$$

- Now let us check that $\text{KDD}(x, y) = 0 \Rightarrow x = y$. We do it by contradiction and thus want to find any pair $x, y, x \neq y$ for which $\text{KDD}(x, y) = 0$.

In particular we need that :

$$\text{KDD}(x, y) = \sqrt{\sum_{\ell} g(\lambda_{\ell})^2 (\mathbf{u}_{\ell}^*[x] - \mathbf{u}_{\ell}^*[y])^2} = 0 \quad (11)$$

with $x \neq y$. Since g is full rank then $g(\lambda_\ell) > 0, \forall \ell$ and thus the only way for (11) to hold is if $\mathbf{u}_\ell^*[x] = \mathbf{u}_\ell^*[y], \forall \ell$. In other words it would imply that the lines x and y of \mathbf{U} are identical. Since \mathbf{U} is a basis, it implies that all its lines are orthonormal, which means there exist no pair x, y such as (11) hold, and thus the contradiction is established, which concludes the proof.

□

Diffusion distance As was hinted in the name, the distance defined in (9) happens to be a generalized diffusion distance. Indeed, taking its spectral formulation we have :

$$d_g(i, j) = \sqrt{\sum_{\ell} g(\lambda_\ell)^2 (\mathbf{u}_\ell^*[i] - \mathbf{u}_\ell^*[j])^2} = D_t(i, j), \quad (12)$$

where $D_t(i, j)$ is the diffusion distance associated to specific kernels depending on t (i.e. the diffusion parameter). If we take two common definitions of the diffusion distance, the original works of [16] and [17] use a kernel of the form $g(x) = x^t$ and the Graph Diffusion Distance defined in [18] uses the heat kernel $g(x) = e^{-tx}$.

5 Graph transductive learning

In this section we want to cast the problem of diffusing the information obtained on a few samples of the data (e.g. using sampling schemes such as defined in Section 3.1) in a transductive inference framework. In this setting, we are observing a label field or signal \mathbf{x} only at a subset of vertices $S \subset V$, i.e $\mathbf{y}_i = \mathbf{x}[i], \forall i \in S$, with \mathbf{y} being the observed signal also called the label function. The goal of transductive learning is to predict the missing signal/labels using both the observed signal and the remaining data points.

5.1 Global graph diffusion

Solutions of transductive inference using graphs can be solved in a number of ways, for example using Tikhonov regression :

$$\arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 + \mu \mathbf{x}^t \mathbf{L} \mathbf{x}, \quad (13)$$

where \mathbf{M} is the sampling operator and \mathbf{L} the graph Laplacian. An alternative to the use of the Dirichlet smoothness constraint is to use graph Total Variation (TV). The regression would thus become :

$$\arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 + \mu \|\nabla_{\mathcal{G}} \mathbf{x}\|_1 \quad (14)$$

with $\nabla_{\mathcal{G}} \mathbf{x} = (\sqrt{\mathbf{W}_{i,j}}(\mathbf{x}[i] - \mathbf{x}[j])), \forall (v_i, v_j) \in \mathcal{E}$.

For large scale learning, solving the optimization problems as described above can be too expensive and one typically uses accelerated descent methods.

5.2 RKHS transductive learning on graphs

5.2.1 Motivation

Our first contribution is to replace the smoothness term arising in 13 by constraining the solution to belong to the finite dimensional Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_{\mathbf{G}}$ corresponding to the graph kernel $\mathbf{G} = g(\mathbf{L})$, for some filter g . In this case, we instead solve the following problem :

$$\arg \min_{\mathbf{x} \in \mathcal{H}_{\mathbf{G}}} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2$$

and show that the solution is given by a simple low-pass filtering step applied to the labelled examples.

5.2.2 Transductive learning and graph filters

In this section, we formulate transductive learning as a finite dimensional regression problem. This problem is solved by constructing a reproducing kernel Hilbert space from a graph filter, which controls the smoothness of the solution and provides a fast algorithm to compute it.

An empirical reproducing kernel Hilbert space Let g be a smooth, strictly positive function defining a graph filter as defined in Section 2. The graph filter defines the following matrix :

$$\mathbf{G}[i, j] = g(\mathbf{L})[i, j] = \mathcal{T}_i g[j],$$

where \mathcal{T}_i is the localisation operator at vertex i . Since the filter is strictly positive definite, \mathbf{G} is positive definite and can be written as the Gram matrix of a set of linearly independent vectors. To see this, we use the spectral representation :

$$\begin{aligned} \mathbf{G} &= \mathbf{U}g(\Lambda)\mathbf{U}^* \\ &= \mathbf{U}g(\Lambda)^{1/2}(\mathbf{U}g(\Lambda)^{1/2})^*. \end{aligned}$$

Let \mathbf{r}_i be the i -th row of $\mathbf{U}g(\Lambda)^{1/2}$, we immediately see that $\mathbf{r}_i^T \mathbf{r}_j = \mathbf{G}[i, j]$. More explicitly, these vectors are written in terms of the graph filter :

$$\mathbf{r}_i[j] = \sum_{\ell} \sqrt{g(\lambda_{\ell})} \mathbf{u}_{\ell}[i] \mathbf{u}_{\ell}[j].$$

These expressions suggest to define the Hilbert space $\mathcal{H}_{\mathbf{G}}$ as the closure of all linear combinations of localized graph filters $\mathcal{T}_i g$. This space is therefore composed of functions of the form :

$$\mathbf{x} = \sum_{k \in \mathcal{V}} \alpha_k \mathcal{T}_k g. \quad (15)$$

Note that any $\mathbf{x} \in \mathcal{H}_{\mathbf{G}}$ has a well-defined graph Fourier transform :

$$\hat{\mathbf{x}}(\ell) = g(\lambda_{\ell}) \sum_{k \in \mathcal{V}} \alpha_k \mathbf{u}_{\ell}[k].$$

This allows to equip $\mathcal{H}_{\mathbf{G}}$ with following scalar product :

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}_{\mathbf{G}}} = \sum_{\ell} \frac{1}{g(\lambda_{\ell})} \hat{\mathbf{x}}(\ell)^* \hat{\mathbf{y}}(\ell)$$

and the vectors \mathbf{r}_i form an orthonormal basis of $\mathcal{H}_{\mathbf{G}}$:

$$\begin{aligned} \langle \mathbf{r}_i, \mathbf{r}_j \rangle_{\mathcal{H}_{\mathbf{G}}} &= \sum_{\ell} \frac{1}{g(\lambda_{\ell})} \sqrt{g(\lambda_{\ell})} \mathbf{u}_{\ell}[i]^* \sqrt{g(\lambda_{\ell})} \mathbf{u}_{\ell}[j] \\ &= \sum_{\ell} \mathbf{u}_{\ell}[i]^* \mathbf{u}_{\ell}[j] \\ &= \delta_{i,j}. \end{aligned}$$

Let us now see that $\mathcal{H}_{\mathbf{G}}$ is a reproducing kernel Hilbert space (rkhs). We show that the scalar product with $\mathcal{T}_i g$ in $\mathcal{H}_{\mathbf{G}}$ is the evaluation functional at vertex i . We first compute :

$$\begin{aligned} \langle \mathcal{T}_i g, \mathcal{T}_j g \rangle_{\mathcal{H}_{\mathbf{G}}} &= \sum_{\ell} \frac{1}{g(\lambda_{\ell})} g(\lambda_{\ell})^2 \mathbf{u}_{\ell}[i]^* \mathbf{u}_{\ell}[j] \\ &= \mathcal{T}_i g[j]. \end{aligned}$$

By linearity of the scalar product and the definition of $\mathcal{H}_{\mathbf{G}}$ (15) we have :

$$\begin{aligned} \langle \mathcal{T}_i g, \mathbf{x} \rangle_{\mathcal{H}_{\mathbf{G}}} &= \sum_{k \in \mathcal{V}} \alpha_k \langle \mathcal{T}_i g, \mathcal{T}_k g \rangle_{\mathcal{H}_{\mathbf{G}}} \\ &= \sum_{k \in \mathcal{V}} \alpha_k \mathcal{T}_k g[i] \\ &= \mathbf{x}[i]. \end{aligned}$$

Finally, for any $\mathbf{x} \in \mathcal{H}_{\mathbf{G}}$, $\mathbf{x} = \sum_{k \in \mathcal{V}} \beta_k \mathcal{T}_k g$, we have the following explicit form of their norm :

$$\begin{aligned}
\|\mathbf{x}\|_{\mathcal{H}_{\mathbf{G}}}^2 &= \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}_{\mathbf{G}}} \\
&= \sum_{\ell} \frac{1}{g(\lambda_{\ell})} g(\lambda_{\ell})^2 \sum_{i,j \in \mathcal{V}} \beta_i \beta_j^* \mathbf{u}_{\ell}[i] \mathbf{u}_{\ell}[j]^* \\
&= \sum_{i,j \in \mathcal{V}} \beta_i \beta_j^* \left(\sum_{\ell} g(\lambda_{\ell}) \mathbf{u}_{\ell}[i] \mathbf{u}_{\ell}[j]^* \right) \\
&= \sum_{i,j \in \mathcal{V}} \beta_i \mathbf{G}[i,j] \beta_j^* \\
&= \beta^T \mathbf{G} \beta.
\end{aligned}$$

Transductive learning Now that we have established $\mathcal{H}_{\mathbf{G}}$ as a valid RKHS, we will seek to recover the full signal by solving the following problem :

$$\tilde{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathcal{H}_{\mathbf{G}}} \sum_{k \in S} L(\mathbf{y}_k, \mathbf{x}[k]) + \mu \|\mathbf{x}\|_{\mathcal{H}_{\mathbf{G}}}. \quad (16)$$

Let us first decompose $\mathcal{H}_{\mathbf{G}} = \mathcal{H}_S \oplus \mathcal{H}_S^{\perp}$, where

$$\mathcal{H}_S = \left\{ \mathbf{x} \in \mathcal{H}_{\mathbf{G}} \text{ s.t. } \mathbf{x} = \sum_{k \in S} \alpha_k \mathcal{T}_k g \right\}.$$

Let us note that, for any $\mathbf{x} \in \mathcal{H}_S$,

$$\begin{aligned}
\|\mathbf{x}\|_{\mathcal{H}_{\mathbf{G}}}^2 &= \sum_{\ell} g(\lambda_{\ell}) \sum_{i,j \in S} \alpha_i \alpha_j^* \mathbf{u}_{\ell}[i] \mathbf{u}_{\ell}[j]^* \\
&= \sum_{i,j \in S} \alpha_i \alpha_j^* \sum_{\ell} g(\lambda_{\ell}) \mathbf{u}_{\ell}[i] \mathbf{u}_{\ell}[j]^* \\
&= \alpha^T \mathbf{K} \alpha
\end{aligned}$$

where $\mathbf{K}[i,j] = \mathbf{G}[i,j]$, $i, j \in S$, is positive definite since it is a principal sub-matrix of a positive definite matrix.

Let $\mathbf{x} \in \mathcal{H}_{\mathbf{G}}$ be decomposed as $\mathbf{x} = \mathbf{x}_S + \mathbf{x}_{S^{\perp}}$, where \mathbf{x}_S (resp. $\mathbf{x}_{S^{\perp}}$) is the orthogonal projection of \mathbf{x} on \mathcal{H}_S (resp. \mathcal{H}_S^{\perp}). Now it is immediate to check that :

$$\begin{aligned}
\langle \mathcal{T}_k g, \mathbf{x}_{S^{\perp}} \rangle_{\mathcal{H}_{\mathbf{G}}} &= \mathbf{x}_{S^{\perp}}[k] \\
&= 0, \forall k \in S.
\end{aligned}$$

Inserting this relationship back into (16), we see that :

$$\sum_{k \in S} L(\mathbf{y}_k, \mathbf{x}_S[k] + \mathbf{x}_{S^{\perp}}[k]) + \lambda \|\mathbf{x}_S + \mathbf{x}_{S^{\perp}}\|_{\mathcal{H}_{\mathbf{G}}}^2 \geq \sum_{k \in S} L(\mathbf{y}_k, \mathbf{x}_S[k]) + \lambda \|\mathbf{x}_S\|_{\mathcal{H}_{\mathbf{G}}}^2,$$

since $\mathbf{x}_{S^\perp}[k] = 0 \ \forall k \in S$ and adding \mathbf{x}_{S^\perp} can only increase the norm of \mathbf{x}_S in \mathcal{H}_G . This shows that the minimizer of (16) is in \mathcal{H}_S and therefore of the form

$$\tilde{\mathbf{x}} = \sum_{k \in S} \beta_k \mathcal{T}_k g$$

for some coefficients β_k . Moreover since $\|\tilde{\mathbf{x}}\|_{\mathcal{H}_G} = \beta^T \mathbf{K} \beta$, we can rewrite (16) as a minimization only on those coefficients with $\tilde{\mathbf{x}} = \mathbf{K} \tilde{\beta}$ and

$$\tilde{\beta} = \arg \min_{\beta} \sum_k L(\mathbf{y}_k, (\mathbf{K} \beta)[k]) + \mu \beta^T \mathbf{K} \beta. \quad (17)$$

Finally, we observe that the recovered signal can be computed by filtering a stream of Kronecker deltas located at the observed values and weighted by the optimal coefficients computed in (17) :

$$\tilde{\mathbf{x}} = g(\mathbf{L}) \left\{ \sum_{k \in S} \tilde{\beta}_k \delta_k \right\}. \quad (18)$$

To summarize, in the case of the squared loss function $L(a, b) = (a - b)^2$, the transductive solution is given by the following two steps algorithm :

1. Compute the optimal coefficients $\tilde{\beta} = (\mathbf{K} + \lambda \mathbb{I})^{-1} \mathbf{y}$
2. Compute the regression $\tilde{\mathbf{x}} = g(\mathbf{L}) \left\{ \sum_{k \in S} \tilde{\beta}_k \delta_k \right\}$.

Note that in traditional ridge regression, the last step is usually given in terms of an explicit kernel that is easy to evaluate. In our case, this expression is also available from (18):

$$\begin{aligned} \tilde{\mathbf{x}}[i] &= \sum_{k \in S} \tilde{\beta}_k \mathbf{G}[i, k] \\ &= \sum_{k \in S} \tilde{\beta}_k \mathbf{G}[k, i] \\ &= \sum_{k \in S} \tilde{\beta}_k \mathcal{T}_k g[i] \end{aligned}$$

and, while the kernel does not have a simple analytical form, the sum can be efficiently computed via a graph filtering algorithm. In particular, it is sufficient to perform $|S|$ filterings to get $\mathcal{T}_k g, \forall k \in S$.

5.3 Convex hull diffusion

If we want to cast the general problem of transductive learning in a simpler framework, we can restrict ourselves to linear solutions of the form $\tilde{\mathbf{x}} = \mathbf{A} \mathbf{y}$. This means finding the coefficients such as :

$$\tilde{\mathbf{x}}[i] = \sum_{k \in S} \alpha_{i,k} \mathbf{y}_k, \quad (19)$$

with $\alpha_{i,k} = \mathbf{A}[i, k]$.

In the previous section, we just saw how a RKHS built on a graph filter g allowed to weight the contributions of localized filters centered on a subset S of vertices. Writing the answer as a linear solution such as defined in (19) would give the following coefficients :

$$\alpha_{i,k} = \frac{\tilde{\beta}_k \mathcal{T}_k g[i]}{\mathbf{y}_k}. \quad (20)$$

Of course, this is kind of a degenerate solution since the coefficients are normalized by \mathbf{y}_k and the optimal coefficients already contain the information from \mathbf{y} .

5.3.1 Convex Hull Diffusion

In this section we propose to use a notion of distances to the samples \mathbf{y} to set the coefficients, more formally $\alpha_{i,k} \simeq d(x_i, y_k)$ for some distance function d . Here, quite naturally, we propose to make use of the LKD as defined in Section 4. Since the coefficients $\alpha_{i,k}$ need to encode similarity between i and k , a reasonable choice is to set :

$$\alpha_{i,k} = 1 - \text{LKD}(i, k) = \frac{\mathcal{T}_i g^2[j]}{\|\mathcal{T}_i g\| \|\mathcal{T}_j g\|}. \quad (21)$$

Using this definition, we know that the coefficients $\alpha_{i,k}$ have good properties derived from Theorem 4. First, since the LKD has values in $[0, 1]$, the coefficients will also have values in this range. Second, $\alpha_{i,k} = \alpha_{k,i}$ which means that \mathbf{A} is symmetric, square and non-negative. Finally, for any kernel g we have $\alpha_{i,i} = 1$ and, if we restrict ourselves to kernels as defined in Theorem 5, we have $\alpha_{i,j} = 0 \Leftrightarrow i = j$. In general, we have the good property that the coefficients $\alpha_{i,k}$ will be small if the vertices i and k are far apart on the graph and big if they are close.

Now, knowing that a classical problem related to embedding data in low dimension, and more specifically to data visualization is a concentration around zero, we wish to devise a method to prevent it. It is reasonable to suppose that the problem of concentration is often related to a lack of information about some points or an absence of normalization. For example, if we take the linear combination as defined in (19), this could happen if for some i , all the coefficients $\alpha_{i,k}$ are small.

In order to avoid this problem, we propose to use a normalized version $\tilde{\mathbf{A}}$ of \mathbf{A} that maps the points \mathbf{x} in the convex hull of \mathbf{y} . This is done simply by normalizing each line of \mathbf{A} , that is :

$$\tilde{\alpha}_{i,k} = \frac{\alpha_{i,k}}{\sum_{k \in S} \alpha_{i,k}} \quad (22)$$

with $\tilde{\alpha}_{i,k} = \tilde{\mathbf{A}}[i, k]$.

6 Compressive Embedding

Building on what has been presented in the previous sections, we now propose our main contribution, a compressive embedding algorithm.

Algorithm 1 is the main algorithm of our proposed scheme. In the following, D denotes the original $N \times K$ data matrix, S the high-dimensional sketch, which is an $M \times K$ subset of D , \mathcal{A}_e is any embedding algorithm, E_S the low-dimensional sketch and E_D an embedding of the full data D being of dimension $M \times d$ and $N \times d$ respectively. \mathcal{D}_G is the diffusion operator on the graph. We have $M < N$, $d < K$ and typically $d = 2$ or $d = 3$ when targeting visualization tasks.

Algorithm 1 Compressive Embedding

- 1: Compute a knn graph \mathcal{G} from the data D
 - 2: Sample M nodes of \mathcal{G} cf. Section 3.1
 - 3: Create a sketch S from D using the sampled nodes
 - 4: Apply \mathcal{A}_e to S to obtain an embedding $E_S = \mathcal{A}_e(S)$
 - 5: Solve the transductive learning problem to get \mathcal{D}_G c.f. Section 5
 - 6: Apply the diffusion operator to obtain the final embedding $E_D = \mathcal{D}_G(E_S)$
-

Let us detail Algorithm 1 step by step.

1. The graph construction can be carried out very efficiently by performing ANN searches in the data. Various methods and optimized libraries are available for this task such as FLANN [9]² or ANNOY³. From our experiments, the graph construction process is not the main computationally intensive task.
2. Guided by the theoretical analysis of Section 3.1 we use low-pass concentrated kernels. Two choices are interesting, either a low-rank approximation (such as defined in Theorem 3) of a heat kernel $g(x) = e^{-\tau x}$ or an exponential window such as $g(x) = s\left(\frac{1-x}{b_{\max}}\right)$ with :

$$s(x) = \begin{cases} 0 & \text{if } x < -1 \\ \frac{e^{-\frac{a}{x}}}{e^{-\frac{a}{x}} + e^{-\frac{a}{1-x}}} & \text{if } x \in [-1, 1] \\ 1 & \text{if } x > 1 \end{cases}$$

where b_{\max} is the desired cut-off frequency.

In Section 3.1 we defined theoretically the number of samples needed to be able to sense and diffuse information from the sampled nodes to every other node. In practice, we were able to verify that $M = \mathcal{O}(\log(N))$, is

²<http://www.cs.ubc.ca/research/flann/>

³<https://github.com/spotify/annoy>

sufficient for the diffusion process. When the number of classes $|\mathcal{C}|$ is available, $M = \mathcal{O}(|\mathcal{C}| \log(N))$ is a good choice. Otherwise $M = \mathcal{O}(d(\mathcal{G}) \log(N))$ is a valid alternative, with $d(\mathcal{G})$ the diameter of the graph. All those choices for M are above the bounds defined in Section 3.1 for any choice of concentration of the kernels since $k < N$.

3. Since there is a trivial mapping between node indices and data points, creating the high-dimensional sketch S is simply taking the subset of D corresponding to the samples indices.
4. The compressive embedding framework does not impose any constraint on the type of algorithm used. Indeed, any embedding algorithm \mathcal{A}_e that can be applied on D , can be applied on $S \subset D$. We note the application of the embedding algorithm $E_S = \mathcal{A}_e(S)$.
5. The proposed transductive learning methods used for the diffusion need only graph filtering operations which are all carried out using Chebyshev polynomial approximations. The two operators that need to be computed are the localized filters $\mathcal{T}_i g$ and $\|\mathcal{T}_i g\|$. The former can be computed by filtering Kronecker delta centered on i , which means that exactly one filtering is needed to compute one $\mathcal{T}_i g$. The 2-norm $\|\mathcal{T}_i g\|$ being needed for all i , one cannot compute it trivially by computing N atoms since it would require N filterings. So instead of computing the exact solution, we can approximate it using random filtering, i.e. $\|\mathcal{T}_i g\|^2$ is well estimated by $\mathbb{E} [\|g(\mathbf{L}) \mathbf{R} \delta_i\|^2]$ with \mathbf{R} an $N \times P$ random matrix. This estimator can be computed by performing only P filterings.
6. The final diffusion is a simple matrix-vector multiplication for both RKHS and CHD methods.

7 Embedding quality measures

In the context of embedding algorithms for visualization two approaches are often used to assess their quality. The first one is a purely qualitative assessment by visual examination, which generally implies to have access to labeled data (see e.g. [5] [6]). When labels are not available, a common practice is to generate the labels using a clustering of the points in high dimension. Visual examination is especially used for relative quality assessment, i.e. one method versus others.

A second method, which is not directly related to visualization, is to measure the quality of the embedding, i.e. if close high dimensional points stay close after embedding. Different numerical measures of local consistency have been proposed such as generalization error of 1-nearest neighbor classifiers [7][19], trustworthiness and continuity [20]. These quantitative assessments do not take into account possible labels for the data.

In order to have quantitative quality measures that take labels into account, we propose three methods that evaluate different characteristics of the embeddings. Note that, despite the fact that we consider the problem settings for which the

data points are associated to some categorical information, data points with no label or multiple labels can be easily accommodated. We will write the set of categorical labels (also called classes) as $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$. For each class c_i we note V_{c_i} the subset of vertices of \mathcal{G}_e having the label c_i .

The common point between all our proposed methods is that they are based on a similarity graph constructed between the points in the embedded domain, that we will call \mathcal{G}_e to distinguish from \mathcal{G} . For simplicity, a simple kNN graph using the Euclidean distance on the embedded points is sufficient. The first method is inspired by Cheeger constants and measures the clusterability of \mathcal{G}_e . The second method uses diffusion distances to measure class homogeneity and the third uses \mathcal{T}_{ig} to estimate the amount of positional outliers.

7.1 Average Clusterability Index

Graph cuts In order to use graph cuts, we start with a few definitions. A cut partitions a graph \mathcal{G} in two complementary sets of vertices S and S^c with $V = S \cup S^c$ and $S \cap S^c = \emptyset$. The graph cut operator is then defined as

$$Cut(S, S^c) = \sum_{i \in S} \sum_{j \in S^c} w_{ij} \quad (23)$$

which represents the total weight of the edges between S and S^c , or the weight of the edges trimmed by the cut.

In order to define the balanced cuts we also need to use the volume operator which is defined as

$$Vol(S) = \sum_{i \in S} d_i \quad (24)$$

where d_i is the degree of the vertex v_i .

Balanced cuts The first interest of cuts in the context of clustering is that the minimization of 23 happens to be a solution to the clustering problem [21]. The minimal cut is however rarely used in practice as it tends to favor small sets of isolated vertices. This led to a shift in focus to balanced cuts, which are cuts normalized by the volume that balances the size of the clusters. Two of the most popular balanced cuts are the Cheeger cut [22] and the Normalized cut [23].

The Cheeger cut is related to the Cheeger constant which is defined as :

$$h(\mathcal{G}) = \min_{S \not\subseteq V} \frac{Cut(S, S^c)}{\min(Vol(S), Vol(S^c))} \quad (25)$$

for a graph \mathcal{G} . This number is a measure of the clusterability of \mathcal{G} , i.e. it is small if there is a strong bottleneck and large otherwise.

Class clusterability The Cheeger cut and cheeger constant imply a minimization in order to find the best clusters, but in our case, we already have the clusters as they are derived from the labels. We can thus reformulate Eq. 25 to define a Cheeger score for a class c_i as :

$$h(\mathcal{G}, c_i) = \frac{Cut(\mathcal{V}_{c_i}, \mathcal{V}_{c_i}^c)}{\min(Vol(\mathcal{V}_{c_i}), Vol(\mathcal{V}_{c_i}^c))} \quad (26)$$

where $\mathcal{V}_{c_i} \subset \mathcal{V}$ is the subset of vertices whose label is c_i and $\mathcal{V}_{c_i}^c \subset \mathcal{V}$ the complementary set containing all the other vertices. We note the number of vertices of a label c_i as $N_{c_i} = |\mathcal{V}_{c_i}|$. Computing the above quantity for a given class give a measure of its clusterability from which we can define the Average Clusterability Index (ACI) as an average weighted by the classes cardinality :

$$ACI = \frac{1}{N} \sum_{c_i \in \mathcal{C}} N_{c_i} h(\mathcal{G}, c_i) = \frac{1}{N} \sum_{c_i \in \mathcal{C}} N_{c_i} \frac{Cut(\mathcal{V}_{c_i}, \mathcal{V}_{c_i}^c)}{\min(Vol(\mathcal{V}_{c_i}), Vol(\mathcal{V}_{c_i}^c))}. \quad (27)$$

This score, as it is inspired by the Cheeger constant, has similar properties : small values mean that the classes are well separated in the graph and large values mean that the classes are much more mixed.

7.2 Average Cluster Concentration

The ACI introduced in the previous section serves to evaluate how clustrable are the different classes. However, this metric will not help discriminate between good clusterability with or without splitted classes. Take for example a dataset with ten classes (such as images of digits). Applying an embedding algorithm could result in having ten classes (the perfect case) or more, meaning that at least one class is splitted in more than one cluster. The ACI between the two cases should be almost indistinguishable, as both embedding scenarii will result in highly clusterable classes.

In order to measure this effect, we need to measure the overall concentration of all points in a class, i.e. that all points in a class are reasonably close to each other. To this end, we introduce a new measure called Average Cluster Concentration which leverages the Kernelized Diffusion Distance introduced above. The principle is that the average distance of all pairs of points of a given class should be small if a class is well concentrated and larger if a class is splitted around different cluster centers.

More formally, using the KDD as defined in 9 and written KDD, we define the ACC for one class $c_i \in \mathcal{C}$ as :

$$\text{ACC}(c_i) = \frac{1}{N_{c_i}^2} \sum_{v_i \in \mathcal{V}_{c_i}} \sum_{v_j \in \mathcal{V}_{c_i}} \text{KDD}(v_i, v_j). \quad (28)$$

As was done above for the ACI, it is natural to give a final score by a weighted average over the classes :

$$\text{ACC} = \frac{1}{N} \sum_{c_i \in \mathcal{C}} N_{c_i} \text{ACC}(c_i) = \frac{1}{N} \sum_{c_i \in \mathcal{C}} \frac{1}{N_{c_i}} \sum_{v_i \in \mathcal{V}_{c_i}} \sum_{v_j \in \mathcal{V}_{c_i}} \text{KDD}(v_i, v_j). \quad (29)$$

This direct computation of the ACC is straightforward but requires $\mathcal{O}(N_{c_i}^2)$ distance evaluations per class. Using the original definition of the KDD, it means making at least $\mathcal{O}(N_{c_i})$ filterings, raising the complexity to $\mathcal{O}(N_{c_i} m |\mathcal{E}|)$ per class assuming order m polynomial approximations for the filtering. Since this is too costly for large graphs, we propose to use a randomized version.

An approach to accelerate the computation of the ACC is to estimate it by randomly picking pairs of points in the class. In order to be robust to different class sizes, we should take a number of samples proportional to N_{c_i} . If we assume that to evaluate n_{c_i} pairs, a reasonable choice is to take $n_{c_i} = \mathcal{O}(N_{c_i})$ which requires a linear number of distance evaluations instead of a quadratic number for the exact ACC computation.

8 Experiments

In this section, we provide experiments whose objective is to show how our proposed methods behave in practice. The first experiments examine how the quantitative measures proposed in Section 7 perform on specially designed synthetic datasets. The second section of experiments allows to visualize the results of the compressive embedding routine using different diffusion operators and compared to state-of-the-art methods.

The experiments were performed with the GSPBox [24], an open-source software. As we stand for reproducible research principles, our implementations and the code to reproduce all our results is open and freely available⁴. Since our methods use random signals, it is expected that the results shall be slightly different in the details, but overall consistent.

8.1 Embedding quality measures

In order to assess the validity of the quantitative measures proposed in Section 7 we use controlled synthetic datasets which exhibit the patterns we would

⁴Will be available online shortly. For now, please contact the corresponding author.

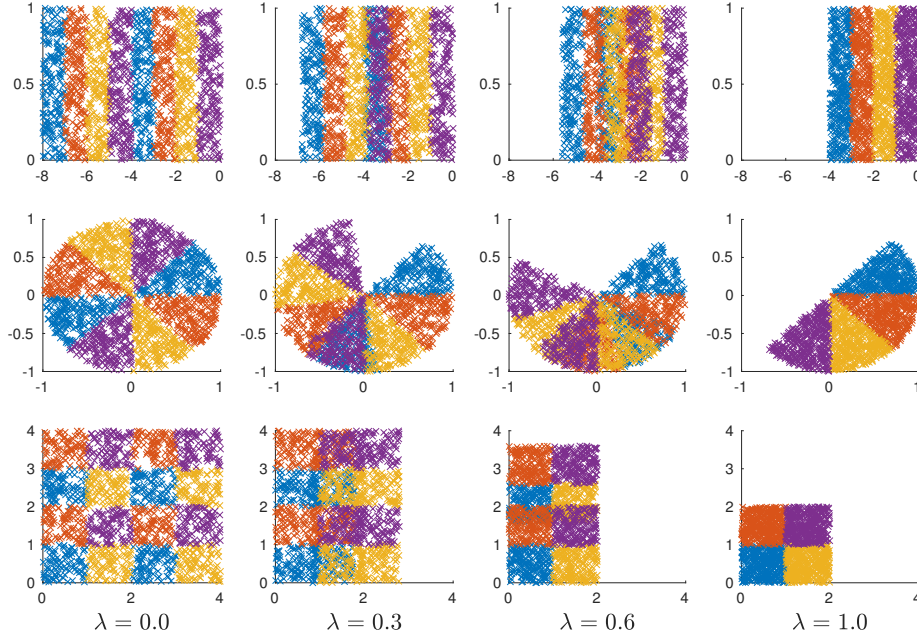


Figure 1: Synthetic datasets with four classes, displayed at dynamics $\lambda = 0, 0.3, 0.6, 1$ (one value per column). On the top row, clusters form bands and move horizontally, on the medium row clusters form disc parts and rotate to form a half-disc and finally on the bottom row clusters are small squares in a larger one, move horizontally until $\lambda = 0.5$ and then vertically.

like to measure. Since we want to evaluate embeddings the datasets are two-dimensional point clouds with labels. All are dynamic and can be deformed continuously between two conformations by varying a parameter $\lambda \in [0, 1]$. Figure 1 displays all datasets for different values of λ .

As can be seen, a unique design principle was used with different topological arrangements. The idea is that for $\lambda = 0$ the different classes are well separated in clusters, with a greater number of clusters than the number of classes. For $\lambda = 1$ the classes are well separated with each class corresponding exactly to one cluster. For intermediate values, the classes are mostly mixed as the points move between the $\lambda = 0$ and $\lambda = 1$ conformations. The checkerboard pattern has an intermediate non-mixed conformation at $\lambda = 0.5$.

Due to the randomness of the data generation process and the evaluation method of the ACC, all results are averages over multiple realisations.

8.1.1 ACI

In this section, we expect to verify that the ACI detects when classes are well clusterized. The results of the ACI scores computed for the three synthetic datasets, using the full dynamic $\lambda \in [0, 1]$ and for different number of classes, is

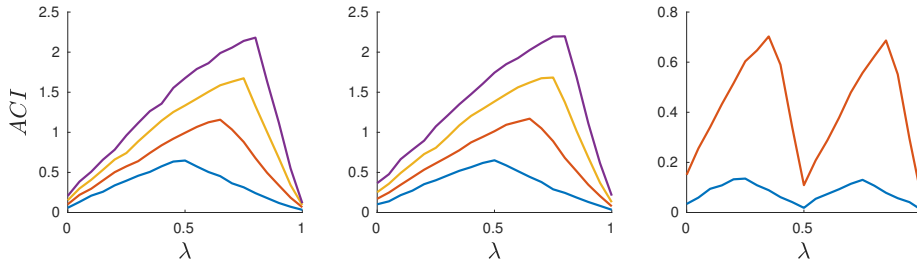


Figure 2: ACI results on synthetic data for the bands (left), circle (middle) and checkerboard (right). The colors indicate the number of classes, blue = 2, orange = 3, yellow = 4 and purple = 5 for left and middle sub-figures, and blue = 4 and orange = 16 on the right.

shown in Figure 2.

As expected, both extreme dynamics ($\lambda = 0$ and $\lambda = 1$ for bands and circle, and additionally $\lambda = 0.5$ for checkerboard) display low ACI scores and the intermediate values correspond to the amount of mixing between the classes. In addition, more classes mean a steeper increase of ACI the the classes mix. As a last remark, we can confirm that the ACI is not sufficient to distinguish between splitted clusters and unified clusters ($\lambda = 0$ and $\lambda = 1$ respectively) which was the main reason for proposing the ACC.

8.1.2 ACC

In this experiment, we want to see if the ACC is able to capture the notion of splitted clusters. Here, the ACC was computed using the randomized method presented in Section 7.2. The results for all datasets can be seen in Figure 2.

The first thing to note is that the curves are not perfectly smooth, due to the randomization process. The general behaviour is however quite clear, for every number of classes. Overall, the results are similar for all datasets and show that the ACC allows to discriminate between $\lambda = 0$ for which we have higher values than for $\lambda = 1$. The result is particularly clear for the bands and checkerboard datasets, and less so for the circle.

8.2 Real-world datasets visualization

In this section, we will present two experiments on real-world datasets for visualization tasks. We restrict ourselves to a relatively small dataset $N < 10^5$ as some of the methods we evaluate cannot scale. We use the classical MNIST⁵ dataset of handwritten digits. It contains 70'000 images of size 28×28 . Note that for this size of dataset the sketch size was 550, which means 0.008% of the data.

⁵<http://yann.lecun.com/exdb/mnist/>

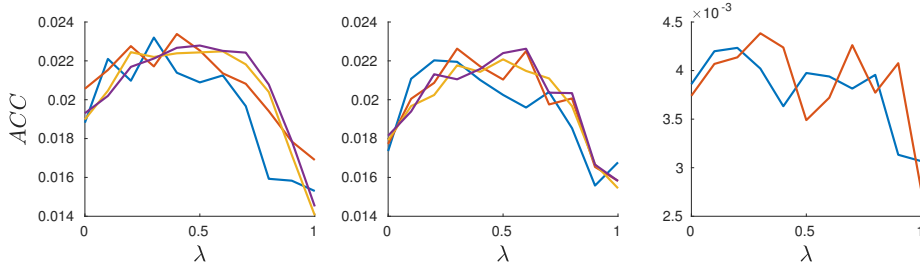


Figure 3: ACC results on synthetic data for the bands (left), circle (middle) and checkerboard (right). The colors indicate the number of classes, blue = 2, orange = 3, yellow = 4 and purple = 5 for left and middle sub-figures, and blue = 4 and orange = 16 on the right.

	Sketch	Tikhonov	RKHS	CHD	Tik+CHD	RKHS+CHD
ACI	2.1035	3.2809	2.3214	1.1054	1.9223	1.6352
ACC	0.0125	0.0312	0.0691	0.0490	0.0491	0.0393

Table 1: ACI and ACC scores for different diffusion operators

8.2.1 Visual comparison of diffusion operators

In this first experiment, we show the resulting embedding of our proposed method using the different graph diffusion operators introduced in Section 5. As a baseline, we also included classical Tikhonov diffusion. Also, in addition to the CHD and RKHS methods, we show the result of bootstrapping Tikhonov and RKHS diffusion with the result of the CHD. The visualizations provided by the 2D embeddings are shown in Figure 4.

Let us begin by inspecting the sketch. The different classes appears to be equally sampled and t-SNE provides a good embedding, while leaving a few overlapping clusters, one splitted class and a few outliers. The Tikhonov and RKHS diffusions achieve a radial separation of the classes but greatly suffer from concentration around zero. The CHD diffusion provides a good embedding similar to the sketch, but tends to produce too much overlaps. The use of bootstrapping as displayed in the last two embeddings seem to improve the results of both Tikhonov and RKHS. By visual inspection, CHD appears to be the best diffusion operator, and in general the convex hull constraint seem to be working as expected.

The quantitative scores for all methods are reported in Table 1. The two worst ACI score are Tikhonov and RKHS, the best one is CHD and the bootstrapped diffusion give medium values. This analysis corresponds well to visual inspection. The ACC scores are very similar and cannot discriminate well between the different methods. This is not surprising since there are no big class splits.

The average timing for the entire process was 161s in total, from which 139s is spent in average on diffusion (step 5 and 6 of Algorithm 1).

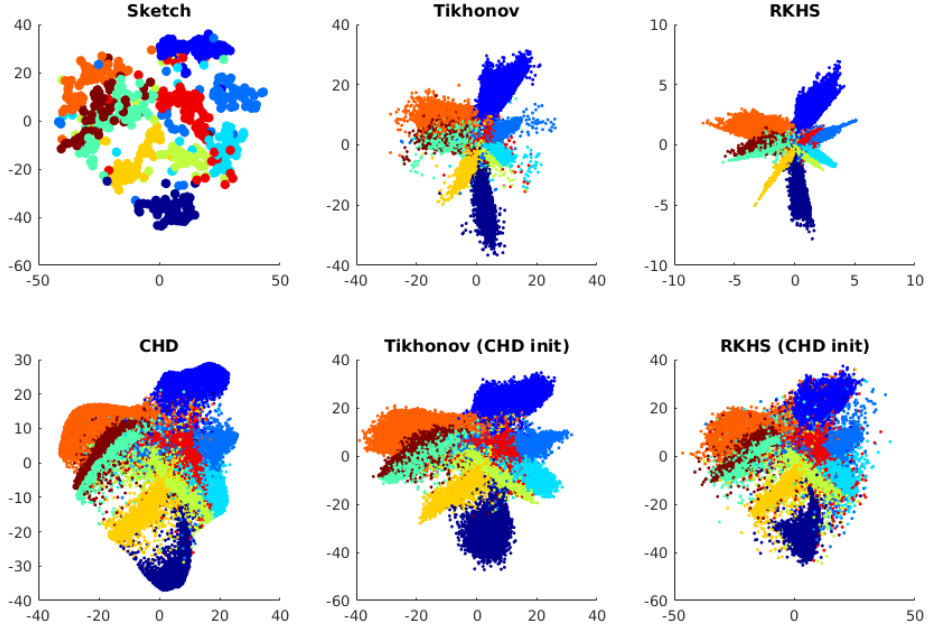


Figure 4: MNIST visualisation using the Compressive Embedding method (with t-SNE as the inner embedding algorithm). The different colors corresponds to the ten different classes.

8.2.2 Original algorithms compared to Compressive Embedding

In this last experiment, we want to see the behaviour of state-of-the-art and traditional visualization algorithm compared to Compressive Embedding versions. We report the visualizations produced, the computing time and the quantitative scores for four different algorithms : t-SNE[5], LargeVis[6], Laplacian Eigenmaps[1] and Sammon mapping[25].

The 2D embeddings produced are shown in Figure 5. If we first look at the original algorithms we can see that both t-SNE and LargeVis produce good embeddings as classes are well separated and clusters are strongly defined. A class split occurs for t-SNE and the repartition is not well balanced for LargeVis but the result is overall very good. Laplacian Eigenmaps gives a fair result but suffers from overlaps and concentration around zero. Sammon Mapping is not shown because the original implementation does not scale enough to complete on a dataset of this size.

Now looking at the sketches we see that both t-SNE and LargeVis produce reasonably good embeddings while leaving a few overlaps, class splits and outliers. Laplacian Eigenmaps suffers from a bit of concentration around zero and tends to mix a few classes together. Sammon Mapping gives a result in which classes are fairly mixed and does not produce well defined clusters.

Finally, the results of the CHD diffusion from the sketches is very consistant accross the different algorithms. Overall CE on t-SNE and LargeVis is quite

ACI	t-SNE	Laplacian Eigenmaps	Sammon Mapping	LargeVis
Original	0.30	2.88	- ¹	0.45
CE	1.98	2.95	3.36	2.19

Table 2: ACI scores comparison between original implementations and Compressive Embedding acceleration.

¹exceeded the maximum memory available (128 GB)

ACC	t-SNE	Laplacian Eigenmaps	Sammon Mapping	LargeVis
Original	0.04	0.04	- ¹	0.03
CE	0.05	0.05	0.04	0.04

Table 3: ACC scores comparison between original implementations and Compressive Embedding acceleration.

¹exceeded the maximum memory available (128 GB)

satisfactory, giving well defined clusters. The downside being too much overlap and a lot of sparse outliers. While being satisfactory, the resulting embeddings are visually less good than their original counterparts. For Laplacian Eigenmaps the CE is very similar to the sketch and difficult to distinguish from its original counterpart. The CE of the Sammon Mapping is surprisingly good given the low quality of the sketch. Visually the result is better after diffusion, as the clusters are reasonably well defined. The problem of overlapping classes and sparse noise is still present.

The ACI and ACC scores for all methods are reported in Table 2 and Table 3. The lowest ACI are for original t-SNE and LargeVis, the second two best results are for CE t-SNE and CE LargeVis. Next, Laplacian Eigenmaps in its original implementation and with CE give similar ACI scores. Finally, Sammon Mapping gives the worst score. All values are very consistent with the visual inspection and tend to validate the use of the ACI as a quantitative measure for embedding quality evaluation. The values reported for the ACC are very similar and do not allow for a very good discrimination since no case of good clustering with class-split was present.

Finally, the computing time is reported in Table 4. For both t-SNE and Laplacian Eigenmaps, CE is one order of magnitude faster than the original implementations. In the case of LargeVis, the CE implementation is still faster but of a smaller factor. However, we need to evaluate this with caution as the original implementation of LargeVis is multi-threaded while all others implementations (including CE) is mono-thread. Taking into account the mono-thread computing time of LargeVis we go back to an order of magnitude acceleration.

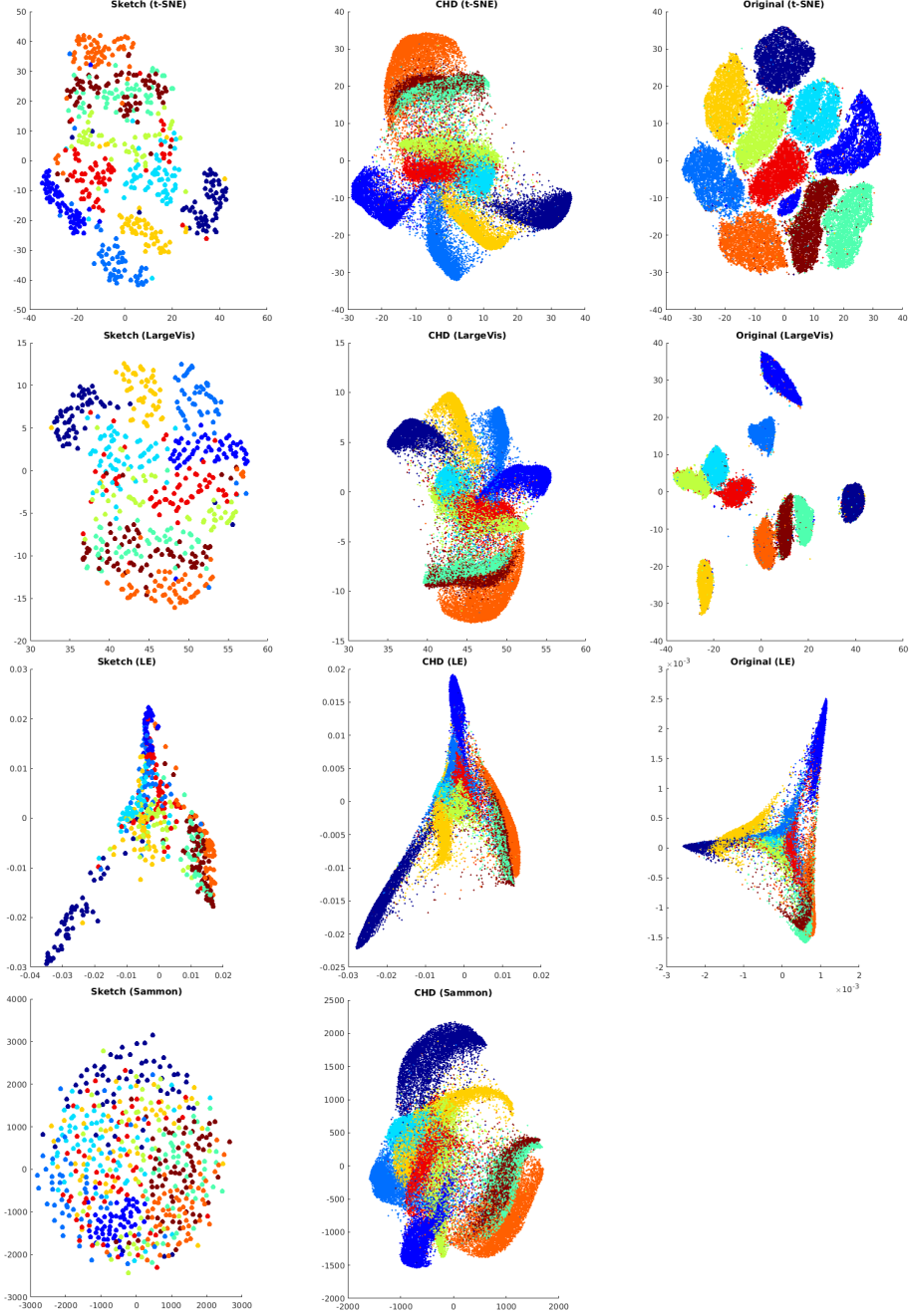


Figure 5: MNIST visualization using different embedding algorithms both in their original implementations (right column) and using Compressive Embedding as an accelerator (middle column). The left column shows the result of the embedding algorithm on the sketch only.

Time [s]	t-SNE	Laplacian Eigenmaps	Sammon Mapping	LargeVis
Original	1815	1666	- ¹	660 ²
CE	157	155	166	329

Table 4: Computing time comparison between original implementations and Compressive Embedding acceleration.

¹exceeded the maximum memory available (128 GB)

²the default implementation uses parallelism, the single thread time usage is 4090s.

9 Conclusion

In this contribution, we have presented a general framework for the acceleration of embedding and visualization algorithms. Our method is made possible by the use of similarity graphs, efficient sampling and graph diffusion. We showed how the method worked on real-world examples and that it gives satisfactory results while being one order of magnitude faster than original implementations. In future works we would like to evaluate active techniques both for sampling and for diffusion.

A Proofs

Important lemmas. Let us first recall two important lemmas necessary for the proofs. The first one is a generalization of the Bernstein inequality for matrices.

Lemma 1 (Matrix Bernstein: Bounded Case). *[26, Theorem 6.1] Consider a finite sequence \mathbf{X}_m of independent, random, self-adjoint matrices with dimension d . Assume that*

$$\mathbb{E}[\mathbf{X}_m] = 0 \quad \text{and} \quad \sigma_{\max}(\mathbf{X}_m) \leq R \quad \text{almost surely.}$$

Compute the norm of the total variance,

$$A^2 := \left\| \sum_m \mathbb{E}[\mathbf{X}_m^2] \right\|_{op}$$

Then the following chain of inequalities holds for all $\delta \geq 0$.

$$\begin{aligned} \mathbb{P} \left[\lambda_{\max} \left(\sum_m \mathbf{X}_m \right) \geq \delta \right] &\leq d \cdot \exp \left(-\frac{A^2}{R^2} \cdot h \left(\frac{R\delta}{A^2} \right) \right) \\ &\leq d \cdot \exp \left(\frac{-\delta^2/2}{A^2 + R\delta/3} \right) \\ &\leq \begin{cases} d \cdot \exp(-3\delta^2/8A^2) & \text{for } \delta \leq A^2/R; \\ d \cdot \exp(-3\delta/8R) & \text{for } \delta \geq A^2/R. \end{cases} \end{aligned}$$

where the function h is defined as $h(u) := (1+u) \log(1+u) - u$ for $u \geq 0$.

The second lemma is a generalization of the triangular inequality for the norm of the localization operator.

Lemma 2. *Given any continuous kernel g and g' , the norm of the localization operator satisfies:*

$$\|\mathcal{T}_i g'\|_2^2 - \|\mathcal{T}_i(|g'| - |g|)\|_2^2 \leq \|\mathcal{T}_i g\|_2^2 \leq \|\mathcal{T}_i g'\|_2^2 + \|\mathcal{T}_i(|g'| - |g|)\|_2^2 \quad (30)$$

Proof. From the definition of the localization operator, we have:

$$\begin{aligned} \|\mathcal{T}_i g\|_2^2 &= \sum_{\ell=0}^{N-1} g^2(\lambda_\ell) \mathbf{u}_\ell^2[i] \\ &= \sum_{\ell=0}^{N-1} (g^2(\lambda_\ell) - g'^2(\lambda_\ell)) \mathbf{u}_\ell^2[i] + \sum_{\ell=0}^{N-1} g'^2(\lambda_\ell) \mathbf{u}_\ell^2[i] \\ &\geq \sum_{\ell=0}^{N-1} (g(\lambda_\ell) - g'(\lambda_\ell))^2 \mathbf{u}_\ell^2[i] + \sum_{\ell=0}^{N-1} g'^2(\lambda_\ell) \mathbf{u}_\ell^2[i] \\ &= \|\mathcal{T}_i g'\|_2^2 + \|\mathcal{T}_i(|g'| - |g|)\|_2^2. \end{aligned} \quad (31)$$

A simple change of variable concludes the proof. The inequality 31 follows from the following assertion. For all λ_ℓ such that $|g(\lambda_\ell)| \leq |g'(\lambda_\ell)|$, we have

$$\begin{aligned} g^2(\lambda_\ell) &= g^2(\lambda_\ell) - g'^2(\lambda_\ell) + g'^2(\lambda_\ell) \\ &= (|g(\lambda_\ell)| - |g'(\lambda_\ell)|)(|g(\lambda_\ell)| + |g'(\lambda_\ell)|) + g'^2(\lambda_\ell) \\ &\geq -(|g'(\lambda_\ell)| - |g(\lambda_\ell)|)(|g'(\lambda_\ell)| - |g(\lambda_\ell)|) + g'^2(\lambda_\ell) \\ &= g'^2(\lambda_\ell) - (|g(\lambda_\ell)| - |g'(\lambda_\ell)|)^2. \end{aligned}$$

For the λ_ℓ such that $|g'(\lambda_\ell)| \leq |g(\lambda_\ell)|$, the inequality $g^2(\lambda_\ell) \geq g'^2(\lambda_\ell) - (|g(\lambda_\ell)| - |g'(\lambda_\ell)|)^2$ is trivially satisfied. \square

Proof of Theorem 1 The proof of Theorem 1 is inspired by [27, Theorem 2] but contains some subtleties.

Proof. Let us define $\boldsymbol{\alpha} = \mathbf{U}_k^* \mathbf{x}$. We first notice that

$$g(\mathbf{L})\mathbf{x} = \mathbf{U}_k \mathbf{U}_k^* \mathbf{x} = \mathbf{U}_k g(\boldsymbol{\Lambda}_k) \boldsymbol{\alpha}$$

The quantity of interest is then rewritten as

$$\begin{aligned} &\frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{-\frac{1}{2}} g(\mathbf{L}) \mathbf{x} \right\|_2^2 - \|g(\mathbf{L}) \mathbf{x}\|_2^2 \\ &= \frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{-\frac{1}{2}} \mathbf{U}_k g(\boldsymbol{\Lambda}_k) \boldsymbol{\alpha} \right\|_2^2 - \|\mathbf{U}_k g(\boldsymbol{\Lambda}_k) \boldsymbol{\alpha}\|_2^2 \\ &= \boldsymbol{\alpha}^* \left(\frac{1}{M} g(\boldsymbol{\Lambda}_k) \mathbf{U}_k^* \mathbf{P}^{-\frac{1}{2}} \mathbf{M}^* \mathbf{M} \mathbf{P}^{-\frac{1}{2}} \mathbf{U}_k g(\boldsymbol{\Lambda}_k) - g(\boldsymbol{\Lambda}_k) g(\boldsymbol{\Lambda}_k) \right) \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^* \mathbf{Y} \boldsymbol{\alpha} \end{aligned}$$

where $\mathbf{Y} = \frac{1}{M}g(\mathbf{\Lambda}_k)\mathbf{U}_k^*\mathbf{P}^{-\frac{1}{2}}\mathbf{M}^*\mathbf{M}\mathbf{P}^{-\frac{1}{2}}\mathbf{U}_kg(\mathbf{\Lambda}_k) - g(\mathbf{\Lambda}_k)g(\mathbf{\Lambda}_k)$. The remaining of the proof focus in characterizing the maximum and the minimum eigenvalue of \mathbf{Y} . To do so, we decompose \mathbf{Y} into a sum of M independent, random, self-adjoint matrices \mathbf{X}_i in order to apply Lemma 1. Let us define

$$\mathbf{X}_i := \frac{1}{M} \left(g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \left(\frac{\boldsymbol{\delta}_{\omega_i}\boldsymbol{\delta}_{\omega_i}^*}{\mathbf{p}_{\omega_i}} - \mathbf{I} \right) \mathbf{U}_kg(\mathbf{\Lambda}_k) \right).$$

It can be verified that

$$\mathbf{Y} = \sum_{i=1}^M \mathbf{X}_i = \sum_{i=1}^M \left(\frac{1}{M}g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \left(\frac{\boldsymbol{\delta}_{\omega_i}\boldsymbol{\delta}_{\omega_i}^*}{\mathbf{p}_{\omega_i}} - \mathbf{I} \right) \mathbf{U}_kg(\mathbf{\Lambda}_k) \right).$$

By construction, the matrices \mathbf{X}_i inherit independence from the random variables $\boldsymbol{\delta}_{\omega_i}$. Furthermore, we have

$$\begin{aligned} \mathbb{E}[\mathbf{X}_i] &= \sum_{n=1}^N p_n \frac{1}{M} \left(g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \left(\frac{\boldsymbol{\delta}_n\boldsymbol{\delta}_n^*}{\mathbf{p}_n} - \mathbf{I} \right) \mathbf{U}_kg(\mathbf{\Lambda}_k) \right) \\ &= \frac{1}{M} \left(g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \left(\sum_{n=1}^N \boldsymbol{\delta}_n\boldsymbol{\delta}_n^* - \mathbf{I} \right) \mathbf{U}_kg(\mathbf{\Lambda}_k) \right) \\ &= \mathbf{0} = \mathbb{E}[-\mathbf{X}_i] \end{aligned}$$

To apply Lemma 1 we need the maximum eigenvalue of \mathbf{X}_i and $-\mathbf{X}_i$.

$$\begin{aligned} \sigma_{\max}(\mathbf{X}_i) &= \sigma_{\max} \left(\frac{1}{M}g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \left(\frac{\boldsymbol{\delta}_{\omega_i}\boldsymbol{\delta}_{\omega_i}^*}{\mathbf{p}_{\omega_i}} - \mathbf{I} \right) \mathbf{U}_kg(\mathbf{\Lambda}_k) \right) \\ &\leq \frac{1}{M} \sigma_{\max} \left(\frac{1}{\mathbf{p}_{\omega_i}} g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \boldsymbol{\delta}_{\omega_i}\boldsymbol{\delta}_{\omega_i}^* \mathbf{U}_kg(\mathbf{\Lambda}_k) \right) \\ &= \frac{1}{M} \sigma_{\max} \left(\frac{1}{\mathbf{p}_{\omega_i}} \boldsymbol{\delta}_{\omega_i}^* \mathbf{U}_kg(\mathbf{\Lambda}_k)g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \boldsymbol{\delta}_{\omega_i} \right) \\ &= \frac{1}{M} \max_i \frac{1}{\mathbf{p}_i} \boldsymbol{\delta}_i^* \mathbf{U}_kg(\mathbf{\Lambda}_k)g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \boldsymbol{\delta}_i \\ &= \frac{1}{M} \max_i \frac{\|\mathcal{T}_i g\|_2^2}{\mathbf{p}_i} \end{aligned}$$

$$\begin{aligned} \sigma_{\max}(-\mathbf{X}_i) &= \sigma_{\max} \left(\frac{1}{M}g(\mathbf{\Lambda}_k)\mathbf{U}_k^* \left(\mathbf{I} - \frac{\boldsymbol{\delta}_{\omega_i}\boldsymbol{\delta}_{\omega_i}^*}{\mathbf{p}_{\omega_i}} \right) \mathbf{U}_kg(\mathbf{\Lambda}_k) \right) \\ &\leq \frac{1}{M} \sigma_{\max}(g^2(\mathbf{\Lambda})) = \frac{1}{M} \|g(\boldsymbol{\lambda})\|_{\infty}^2 \end{aligned}$$

Finally, before we can apply Lemma 1, we need to compute

$$\begin{aligned}
A^2 &= \sigma_{\max} \left(\mathbb{E} \left[\sum_{i=1}^M \mathbf{X}_i^2 \right] \right) \\
&= \sigma_{\max} \left(\mathbb{E} \left[\frac{1}{M^2} \sum_{i=1}^M g(\Lambda_k) \mathbf{U}_k^* \left(\mathbf{I} - \frac{\delta_{\omega_i} \delta_{\omega_i}^*}{\mathbf{p}_{\omega_i}} \right) \mathbf{U}_k g(\Lambda_k) g(\Lambda_k) \mathbf{U}_k^* \left(\mathbf{I} - \frac{\delta_{\omega_i} \delta_{\omega_i}^*}{\mathbf{p}_{\omega_i}} \right) \mathbf{U}_k g(\Lambda_k) \right] \right) \\
&= \frac{1}{M} \sigma_{\max} \left(g(\Lambda_k) \mathbf{U}_k^* \mathbb{E} \left[\left(\mathbf{I} - \frac{\delta_{\omega_i} \delta_{\omega_i}^*}{\mathbf{p}_{\omega_i}} \right) \mathbf{U}_k g(\Lambda_k) g(\Lambda_k) \mathbf{U}_k^* \left(\mathbf{I} - \frac{\delta_{\omega_i} \delta_{\omega_i}^*}{\mathbf{p}_{\omega_i}} \right) \right] \mathbf{U}_k g(\Lambda_k) \right) \\
&= \frac{1}{M} \sigma_{\max} \left(g(\Lambda_k) \mathbf{U}_k^* \left(\sum_{i=1}^N \frac{\|\mathcal{T}_i g\|_2^2}{\mathbf{p}_i} \delta_i \delta_i^* \right) \mathbf{U}_k g(\Lambda_k) \right) \\
&\leq \frac{1}{M} \|g(\lambda)\|_\infty^2 \max_i \frac{\|\mathcal{T}_i g\|_2^2}{\mathbf{p}_i},
\end{aligned}$$

since

$$\begin{aligned}
&\mathbb{E} \left[\left(\mathbf{I} - \frac{\delta_{\omega_i} \delta_{\omega_i}^*}{\mathbf{p}_{\omega_i}} \right) \mathbf{U}_k g(\Lambda_k) g(\Lambda_k) \mathbf{U}_k^* \left(\mathbf{I} - \frac{\delta_{\omega_i} \delta_{\omega_i}^*}{\mathbf{p}_{\omega_i}} \right) \right] \\
&= \sum_{i=1}^N \mathbf{p}_i \left(\mathbf{I} - \frac{\delta_i \delta_i^*}{\mathbf{p}_i} \right) \mathbf{U}_k g^2(\Lambda_k) \mathbf{U}_k^* \left(\mathbf{I} - \frac{\delta_i \delta_i^*}{\mathbf{p}_i} \right) \\
&= \sum_{i=1}^N \mathbf{p}_i \delta_i \delta_i^* g^2(L) \delta_i \delta_i^* - g^2(L) \\
&= \sum_{i=1}^N \frac{\|\mathcal{T}_i g\|_2^2}{\mathbf{p}_i} \delta_i \delta_i^* - g^2(L) \\
&\preceq \sum_{i=1}^N \frac{\|\mathcal{T}_i g\|_2^2}{\mathbf{p}_i} \delta_i \delta_i^T
\end{aligned}$$

Let us denote $\max_i \frac{\|\mathcal{T}_i g\|_2^2}{\mathbf{p}_i} = \alpha$. We now apply Lemma 1 to the $\mathbf{Y} = \sum_{i=1}^M \mathbf{X}_i$ and we find

$$\mathbb{P} \left[\frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{-\frac{1}{2}} g(\mathbf{L}) \mathbf{x} \right\|_2^2 - \|g(\mathbf{L}) \mathbf{x}\|_2^2 \geq \delta \|\alpha\|_2^2 \right] \leq k \exp \left(- \frac{M \frac{\delta^2}{2}}{\alpha \left(\|g(\lambda)\|_\infty^2 + \frac{\delta}{3} \right)} \right).$$

Similarly for $-\mathbf{Y} = \sum_{i=1}^M -\mathbf{X}_i$, we find

$$\mathbb{P} \left[\|g(\mathbf{L}) \mathbf{x}\|_2^2 - \frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{-\frac{1}{2}} g(\mathbf{L}) \mathbf{x} \right\|_2^2 \geq \delta \|\alpha\|_2^2 \right] \leq k \exp \left(- \frac{M \frac{\delta^2}{2}}{\|g(\lambda)\|_\infty^2 (\alpha + \frac{\delta}{3})} \right).$$

In order to optimize the bound, we need to minimize α . Thus we choose $\mathbf{p}_i = \frac{\|\mathcal{T}_i g\|_2^2}{\|g(\lambda)\|_2^2}$ and we get $\alpha = \|g(\lambda)\|_2^2$. The two previous inequalities become

$$\mathbb{P} \left[\frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{-\frac{1}{2}} g(\mathbf{L}) \mathbf{x} \right\|_2^2 - \|g(\mathbf{L}) \mathbf{x}\|_2^2 \geq \delta \|\mathbf{U}_k^* \mathbf{x}\|_2^2 \right] \leq k \exp \left(- \frac{M \delta^2}{2 \|g(\lambda)\|_2^2 \left(\|g(\lambda)\|_\infty^2 + \frac{\delta}{3} \right)} \right)$$

$$\mathbb{P} \left[\|g(\mathbf{L})\mathbf{x}\|_2^2 - \frac{1}{M} \left\| \mathbf{M}\mathbf{P}^{-\frac{1}{2}}g(\mathbf{L})\mathbf{x} \right\|_2^2 \geq \delta \|\mathbf{U}_k^*\mathbf{x}\|_2^2 \right] \leq k \exp \left(-\frac{M\delta^2}{2\|g(\boldsymbol{\lambda})\|_\infty^2 \left(\|g(\boldsymbol{\lambda})\|_2^2 + \frac{\delta}{3} \right)} \right)$$

We make the change of variables $\delta' \|g(\boldsymbol{\lambda})\|_\infty^2 = \delta$

$$\begin{aligned} \mathbb{P} \left[\frac{\frac{1}{M} \left\| \mathbf{M}\mathbf{P}^{-\frac{1}{2}}g(\mathbf{L})\mathbf{x} \right\|_2^2 - \|g(\mathbf{L})\mathbf{x}\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} \geq \delta' \|\mathbf{U}_k^*\mathbf{x}\|_2^2 \right] &\leq k \exp \left(-\frac{1}{2} \frac{\|g(\boldsymbol{\lambda})\|_\infty^2}{\|g(\boldsymbol{\lambda})\|_2^2} \frac{M\delta'^2}{\left(1 + \frac{\delta'}{3}\right)} \right) \\ \mathbb{P} \left[\frac{\|g(\mathbf{L})\mathbf{x}\|_2^2 - \frac{1}{M} \left\| \mathbf{M}\mathbf{P}^{-\frac{1}{2}}g(\mathbf{L})\mathbf{x} \right\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} \geq \delta \|\mathbf{U}_k^*\mathbf{x}\|_2^2 \right] &\leq k \exp \left(-\frac{1}{2} \frac{M\delta'^2}{\left(\frac{\|g(\boldsymbol{\lambda})\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} + \frac{\delta'}{3} \right)} \right) \\ &\leq k \exp \left(-\frac{1}{2} \frac{\|g(\boldsymbol{\lambda})\|_\infty^2}{\|g(\boldsymbol{\lambda})\|_2^2} \frac{M\delta'^2}{\left(1 + \frac{\delta'}{3}\right)} \right) \end{aligned} \quad (32)$$

Finally, we substitute δ for δ' . We set the success probability of the event

$$\left| \frac{\frac{1}{m} \left\| \mathbf{M}\mathbf{P}^{-\frac{1}{2}}\mathbf{U}g(\boldsymbol{\Lambda})\mathbf{x} \right\|_2^2 - \|\mathbf{U}g(\boldsymbol{\Lambda})\mathbf{x}\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} \right| \geq \delta \|\mathbf{x}\|_2^2.$$

to $1 - \epsilon$. As both sides of the bound have to be taken into account, we need

$$\frac{\epsilon}{2} \geq k \exp \left(-\frac{1}{2} \frac{\|g(\boldsymbol{\lambda})\|_\infty^2}{\|g(\boldsymbol{\lambda})\|_2^2} \frac{Mt^2}{\left(1 + \frac{\delta}{3}\right)} \right),$$

which is equivalent to impose on M

$$M \geq 2 \frac{1}{\delta^2} \frac{\|g(\boldsymbol{\lambda})\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} \left(1 + \frac{\delta}{3}\right) \log \left(\frac{2k}{\epsilon} \right)$$

□

Proof of Theorem 2

Proof. Given $M \geq 2 \frac{1}{\delta^2} \frac{\|g(\boldsymbol{\lambda})\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} \left(1 + \frac{\delta}{3}\right) \log \left(\frac{k}{\epsilon} \right)$, we use (32) and set $\mathbf{x} = \boldsymbol{\delta}_i$. Then with a probability ϵ , we have

$$\frac{\|\mathcal{T}_i g\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} - \frac{\frac{1}{M} \left\| \mathbf{M}\mathbf{P}^{-\frac{1}{2}}\mathcal{T}_i g \right\|_2^2}{\|g(\boldsymbol{\lambda})\|_\infty^2} \geq \delta \|\mathbf{U}_k^*\boldsymbol{\delta}_i\|_2^2.$$

As a result, with a probability $1 - \epsilon$, we have

$$\frac{\frac{1}{M} \left\| \mathbf{M}\mathbf{P}^{-\frac{1}{2}}\mathcal{T}_i g \right\|_2^2}{\|\mathcal{T}_i g\|_2^2} \geq 1 - \delta \frac{\|g(\boldsymbol{\lambda})\|_\infty^2 \|\mathbf{U}_k^*\boldsymbol{\delta}_i\|_2^2}{\|\mathcal{T}_i g\|_2^2}.$$

The change of variable $\delta' = \delta \frac{\|g(\boldsymbol{\lambda})\|_\infty^2 \|\mathbf{U}_k^*\boldsymbol{\delta}_i\|_2^2}{\|\mathcal{T}_i g\|_2^2}$ concludes the proof. For the factor $\frac{\delta}{3}$, we use the fact that $\frac{\|g(\boldsymbol{\lambda})\|_\infty^2 \|\mathbf{U}_k^*\boldsymbol{\delta}_i\|_2^2}{\|\mathcal{T}_i g\|_2^2} \geq 1$. □

Proof of Theorem 3

Proof. We first use the fact that $\|\mathbf{A}\mathcal{T}_i g'\|_2 \geq \|\mathbf{A}\mathcal{T}_i g\|_2$ for any linear operator \mathbf{A} . This comes from the fact that $\mathcal{T}_i g$ for a fixed i can be written as $\mathbf{T}_i g(\boldsymbol{\lambda})$ where \mathbf{T}_i is a linear operator. We successively apply Theorem 2 and in similar way to Theorem 2, Equation 32 to obtain

$$\begin{aligned} \frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{\frac{1}{2}} \mathcal{T}_i g \right\|_2^2 &\geq \frac{1}{M} \left\| \mathbf{M} \mathbf{P}^{\frac{1}{2}} \mathcal{T}_i g' \right\|_2^2 \\ &\geq \left\| \mathcal{T}_i g' \right\|_2^2 - \delta \left\| g'(\boldsymbol{\lambda}) \right\|_\infty^2 \left\| \mathbf{U}_k^* \boldsymbol{\delta}_i \right\|_2^2 \\ &\geq \left\| \mathcal{T}_i g \right\|_2^2 - \left\| \mathcal{T}_i (|g'| - |g|) \right\|_2^2 - \delta \left\| g'(\boldsymbol{\lambda}) \right\|_\infty^2 \left\| \mathbf{U}_k^* \boldsymbol{\delta}_i \right\|_2^2, \end{aligned}$$

for a number of samples

$$M \geq 2 \frac{1}{\delta^2} \frac{\left\| g'(\boldsymbol{\lambda}) \right\|_2^2 \left\| g'(\boldsymbol{\lambda}) \right\|_\infty^2 \left\| \mathbf{U}_k^* \boldsymbol{\delta}_i \right\|_2^4}{\left\| \mathcal{T}_i g' \right\|_2^4} \left(1 + \frac{\delta}{3} \right) \log \left(\frac{k}{\epsilon} \right).$$

The change of variable $\delta' = \delta \frac{\left\| g'(\boldsymbol{\lambda}) \right\|_\infty^2 \left\| \mathbf{U}_k^* \boldsymbol{\delta}_i \right\|_2^2}{\left\| \mathcal{T}_i g \right\|_2^2}$ and the division by $\left\| \mathcal{T}_i g \right\|_2^2$ conclude the proof. For the factor $\frac{\delta}{3}$, we use the fact that $\frac{\left\| g'(\boldsymbol{\lambda}) \right\|_\infty^2 \left\| \mathbf{U}_k^* \boldsymbol{\delta}_i \right\|_2^2}{\left\| \mathcal{T}_i g \right\|_2^2} \geq 1$. \square

Acknowledgment

We would like to thank Lionel Martin for valuable discussions.

References

- [1] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.
- [2] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [3] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [4] G. E. Hinton and S. T. Roweis, “Stochastic neighbor embedding,” in *Advances in neural information processing systems*, pp. 833–840, 2002.
- [5] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [6] J. Tang, J. Liu, M. Zhang, and Q. Mei, “Visualizing large-scale and high-dimensional data,” in *Proceedings of the 25th International Conference on World Wide Web*, pp. 287–297, International World Wide Web Conferences Steering Committee, 2016.

- [7] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, pp. 66–71, 2009.
- [8] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms.," *Journal of machine learning research*, vol. 15, no. 1, pp. 3221–3245, 2014.
- [9] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, 2014.
- [10] F. R. Chung, *Spectral graph theory*, vol. 92. AMS Bookstore, 1997.
- [11] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "Vertex-frequency analysis on graphs," *arXiv preprint arXiv:1307.5708*, 2013.
- [12] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.
- [13] A. Susnjara, N. Perraudin, D. Kressner, and P. Vandergheynst, "Accelerated filtering on graphs using lanczos method," *arXiv preprint arXiv:1509.04537*, 2015.
- [14] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "Vertex-frequency analysis on graphs," *Applied and Computational Harmonic Analysis*, vol. 40, no. 2, pp. 260–291, 2016.
- [15] G. Puy, N. Tremblay, R. Gribonval, and P. Vandergheynst, "Random sampling of bandlimited signals on graphs," *Applied and Computational Harmonic Analysis*, 2016.
- [16] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators," *arXiv preprint math/0506090*, 2005.
- [17] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.
- [18] D. K. Hammond, Y. Gur, and C. R. Johnson, "Graph diffusion distance: A difference measure for weighted graphs based on the graph laplacian exponential kernel," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pp. 419–422, IEEE, 2013.
- [19] G. Sanguinetti, "Dimensionality reduction of clustered data sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 535–540, 2008.
- [20] J. Venna and S. Kaski, "Visualizing gene interaction graphs with local multidimensional scaling.," in *ESANN*, vol. 6, pp. 557–562, 2006.
- [21] Z. Wu and R. Leahy, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 15, no. 11, pp. 1101–1113, 1993.

- [22] J. Cheeger, “A lower bound for the smallest eigenvalue of the laplacian,” 1969.
- [23] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [24] N. Perraudin, J. Paratte, D. Shuman, L. Martin, V. Kalofolias, P. Vandergheynst, and D. K. Hammond, “GSPBOX: A toolbox for signal processing on graphs,” *ArXiv e-prints*, Aug. 2014.
- [25] J. W. Sammon, “A nonlinear mapping for data structure analysis,” *IEEE Transactions on computers*, vol. 100, no. 5, pp. 401–409, 1969.
- [26] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.
- [27] A. E. Alaoui and M. W. Mahoney, “Fast randomized kernel ridge regression with statistical guarantees,” pp. 775–783, 2015.